ORSAY
N° d'ordre : **2960**

THÈSE

présentée à

L'UNIVERSITÉ PARIS-SUD
CENTRE D'ORSAY

pour obtenir

LE DIPLOME DE DOCTEUR DE 3ème CYCLE

par

Philippe RICHARD

Spécialité : **Informatique**

Sujet de la thèse :

SUR L'ÉVALUATION DE LA TAILLE DE LA RÉPONSE À UNE REQUÊTE DANS UNE BASE DE DONNÉES RELATIONNELLE

Soutenue le 10 avril 1981 devant la Commission composée de :

MM. **GELENBE** Président
BANCILHON
DEMOLOMBE
GARDARIN
ARMSTRONG

THÈSE

présentée à

L'UNIVERSITÉ PARIS-SUD
CENTRE D'ORSAY

pour obtenir

LE DIPLOME DE DOCTEUR DE 3ème CYCLE

par

Philippe RICHARD

Spécialité : **Informatique**

Sujet de la thèse :

SUR L'ÉVALUATION DE LA TAILLE DE LA RÉPONSE À UNE REQUÊTE DANS UNE BASE DE DONNÉES RELATIONNELLE

Soutenue le 10 avril 1981 devant la Commission composée de :

MM. **GELENBE** Président
BANCILHON
DEMOLOMBE
GARDARIN
ARMSTRONG

UNIV.
LEIDEN
BIBL.

THESE DE 3ème CYCLE

SUR L'EVALUATION DE LA TAILLE DE LA REPONSE A UNE REQUETE DANS UNE BASE DE DONNEES RELATIONNELLE

Philippe RICHARD

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41547321_0021

TABLE DES MATIERES

Résumé

CHAPITRE I - "MODELE"

I - Introduction
II - Le Problème de l'Optimisation
III - Définition et Notations
IV - Interrogation de la Base de Données et Définitions Préliminaires
V - Le Modèle

CHAPITRE II - "RESULTATS"

I - Les Classes de Schémas Probabilisés
II - Les Systèmes de Paramètres (Définition)
III - Validité des Systèmes de Paramètres

CHAPITRE III - "LIENS AVEC LES CONTRAINTES D'INTEGRITE"

I - Relations entre les Systèmes de Paramètres et les Dépendances Fonctionnelles
II - Liens avec les Dépendances Multivaluées et les Décompositions du Premier Ordre

CHAPITRE IV - "FERMETURE DES CLASSES"

I - Fermeture de la Classe CE pour les Opérations de l'Algèbre Relationnelle

II - Fermeture de la Classe CI pour les Opérations de l'Algèbre Relationnelle

CHAPITRE V - "VALIDITE DES SYSTEMES P, P_A ET P"

I - Validité des Systèmes P et P_A dans la Classe CE pour les Opérations de l'Algèbre Relationnelle

II - Validité du Système P" dans la Classe CE pour les Opérations de l'Algèbre Relationnelle.

CHAPITRE VI - "VALIDITE DU SYSTEME P'

I - Définitions (Rappel)

II - Validité du Système P' dans la Classe CI

CONCLUSION

1) - Modèle Formel

2) - Applications

Références

Je tiens à remercier,

les membres de mon jury :

Monsieur le Professeur Gelenbe pour m'avoir fait l'honneur de le présider,

Monsieur le Professeur Armstrong pour ses remarques qui m'ont aidé à mieux formaliser ce travail,

Monsieur le Professeur Gardarin pour m'avoir fait l'honneur de faire partie de mon jury,

Monsieur Demolombe, Ingénieur de Recherche au CERT, dont les appréciations ont été très stimulantes et ont permis de fructueuses discussions,

et enfin, Monsieur le Professeur Bancilhon qui m'a dirigé avec gentillesse, tout au long de ce travail et avec qui j'ai beaucoup appris sur le "relationnel" en particulier, et la recherche en général...

ainsi que :

Madame C. Decalf et Mademoiselle I. Bourgeois pour leur magnifique travail de frappe et leur bonne humeur qui a résisté à mon graphisme.

RESUME

Nous nous attachons ici à déterminer les moyens nécessaires à l'évaluation de la taille de la réponse à une question dans un système de gestion de base de données relationnel. En effet, cette évaluation est un outil nécessaire à l'optimisation du traitement d'une requête. Nous définissons le cadre formel de cette étude qui se place dans le cas où le langage spécifiant les requêtes est l'algèbre relationnelle et nous construisons le modèle probabiliste nécessaire à la formalisation du problème.

Nous définissons formellement la notion de validité d'un système de paramètres. Une requête est composée d'une suite d'opérateurs élémentaires de l'algèbre relationnelle appliqués sur les relations de la base et sur les relations, dites dérivées, obtenues par l'application d'opérateurs précédents. Nous nous intéressons à des classes de bases de données caractérisées par des propriétés particulières de leurs relations (indépendance etc...).

Ces propriétés ainsi que les paramètres définis sur la base de données permettent d'évaluer la distribution ou l'espérance de la taille d'une relation obtenue en appliquant un opérateur élémentaire de l'algèbre sur la base de données.

Pour évaluer la taille de la réponse à une requête complexe ce résultat ne suffit pas. En effet, il faut garantir que la base de données "augmentée" de la relation obtenue par l'application d'un opérateur appartient toujours à la classe d'origine. Ainsi lors de l'évaluation de la relation dérivée obtenue en appliquant un opérateur sur la nouvelle base, nous pouvons utiliser les propriétés de

la base. C'est ce que nous appelons fermeture d'une classe de base de données pour les opérations de l'algèbre. De plus, le système de paramètres doit nous permettre non seulement d'évaluer la distribution ou l'espérance de la taille d'une relation construite par un opérateur de l'algèbre mais aussi de calculer les nouveaux paramètres sur cette relation à partir des anciens. Alors le système est dit valide dans une classe de bases de données pour les opérateurs de l'algèbre relationnelle.

Nous montrons que ces propriétés des systèmes de paramètres et des classes de base de données permettent d'évaluer la taille de toute requête construite à partir de l'algèbre. Nous définissons deux classes particulières de bases de données et nous montrons qu'il existe quatre systèmes de paramètres valides dans ces classes pour l'algèbre.

Nos deux classes sont caractérisées par des propriétés d'indépendance entre relations (la présence d'un n-uplet dans une relation est indépendante de sa présence dans l'autre) et d'indépendance à l'intérieur de chaque relation (la présence d'un n-uplet dans une relation donnée est indépendante de la présence d'un quelconque autre n-uplet). De plus l'une des deux classes (notée CE) est caractérisée par une propriété d'équiprobabilité d'appartenance des n-uplets à une relation.

Le premier de nos systèmes de paramètres (noté P) est constitué des espérances des tailles des projections des relations de la base.

Le second (P_A) contient les espérances des tailles des projections uniquement sur les attributs simples ainsi que les tailles des relations. Ces deux systèmes permettent d'évaluer l'espérance de la taille de la réponse à une requête.

Le troisième (P') contient les probabilités d'appartenance d'un n-uplet à une relation et le dernier (P") est composé des distributions de taille des projections de relations. Ces deux derniers

systèmes permettent d'évaluer la distribution de taille des réponses aux requêtes.

Ces systèmes nécessitent une quantité d'information différente et donnent des résultats d'évaluation plus ou moins fins.
Une relation d'ordre est définie qui permet de comparer la finesse de l'évaluation qu'ils permettent. Nous montrons enfin que la connaissance des contraintes d'intégrité caractérisant la base de données peut être utilisée d'une part afin de diminuer la complexité des paramètres nécessaires et d'autre part de simplifier les calculs conduisant à l'évaluation des tailles des relations dérivées. Ainsi la famille de dépendances fonctionnelles de la base permet de construire des systèmes de paramètres équivalents à ceux définis auparavant mais nécessitant une moins grande quantité d'information. De même la connaissance des dépendances multi-valuées ou des dépendances hiérarchiques du premier ordre, permet aussi de réduire la quantité d'information nécessaire et facilite les calculs d'évaluation.

CHAPITRE I

- Modèle -

I - Introduction

II - Le problème de l'optimisation

III - Définitions et notations

IV - Interrogation de la base de données et définitions préliminaires

V - Le modèle

I - INTRODUCTION

L'approche relationnelle dans les systèmes de gestion de base de données autorise un fort degré d'indépendance des données par l'utilisation de langages de requête de haut niveau. En effet, le modèle relationnel permet à l'utilisateur de ne spécifier que le résultat désiré dans la requête. Aucun chemin d'accès aux données n'est indiqué contrairement aux deux autres modèles (hiérarchique et réseau) qui eux, impliquent la connaissance de ces chemins d'accès. Il est donc très important de traiter les requêtes de manière efficace. L'optimisation du traitement des requêtes doit donc comprendre le choix des meilleurs chemins d'accès et la détermination de l'ordre optimal dans lequel les opérateurs relationnels seront traités. Pour faire ces deux choix, il est nécessaire que le processeur d'optimisation possède des moyens lui permettant d'évaluer à chaque étape la taille des relations de base et des relations dérivées qui sont considérées lors du traitement de la requête.

Nous allons illustrer par un exemple l'importance de cette évaluation.

Une chaine de magasins possède une base de données répartie de la façon suivante :

A Toulouse sont présentes les relations :
emp T (nom,rayon,sal T)
dir T (rayon, dir T)

A Marseille : emp M (nom,rayon,sal M)
dir M (rayon,dir M)

A Lille : emp L (nom,rayon,sal L)
dir L (rayon,dir L)

A Paris : emp P (nom,rayon,sal P)
dir P (rayon,dir P)
dir G (rayon,dir,ville)

où emp x caractérise les employés de la succursale de la ville x, et dir x caractérise les directeurs de rayon de la ville x. dir G est un annuaire de tous les directeurs de rayon pour toute la France.

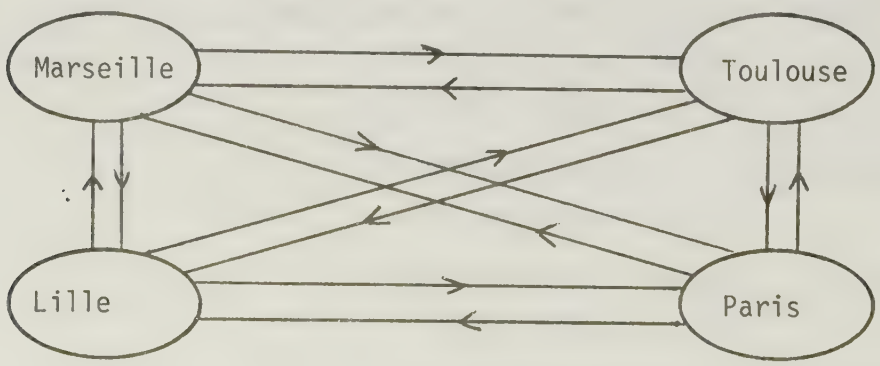
Pour des raisons de clarté, dans les relations emp x et dir x les attributs "sal x" et "dir x" sont personnalisés afin de ne pas créer d'ambiguité dans l'élaboration de la requête citée en exemple. Par exemple, la relation emp T pourra être :

emp T	nom	rayon	sal T
	toto	jouet	10000
	lulu	chaussure	15000

et la relation emp M sera :

emp M	nom	rayon	sal M
	jojo	jouet	11000
	jacques	chaussure	12000

si nous effectuons une opération entre emp T et emp M, il sera créé deux colonnes différentes pour sal T et sal M et une seule pour nom ou rayon. (Nous supposons qu'il n'y a pas de directeurs différents portant le même nom). La base est organisée de la façon suivante :



l'effet de cette organisation est qu'il n'y a pas de site privilégié et que les informations circulent d'un site à un autre directement.

Un utilisateur situé à Paris pose la requête suivante :

"Nom et salaire des employés de Toulouse dont le directeur de rayon gagne plus que le directeur du même rayon à Marseille et à Lille. La réponse doit être donnée à Paris ?"
En algèbre relationnelle, la requête peut se formuler ainsi :

$$R_1 = \text{emp T [rayon = rayon] dir T}$$

R_1 représente les employés de Toulouse, leur salaire, leur rayon et leur directeur.

$$R_1' = R_1 \mid \text{nom = dir}$$

R_1' représente les informations sur les salaires des directeurs de rayon de Toulouse.

$$R_2 = \text{emp M [rayon = rayon] dir M}$$

$$R_2' = R_2 \mid \text{nom = dir}$$

R_2 et R_2' sont les homologues de R_1 et R_1' pour Marseille.

$$R_3 = \text{emp L [rayon = rayon] dir L}$$

$$R_3' = R_3 \mid \text{nom = dir}$$

R_3 et R_3' sont les homologues de R_1 et R_1' pour Lille.

$$R_{12}' = R_1' \text{ [rayon = rayon] } R_2'$$

R_{12}' associe les directeurs de Toulouse et Marseille travaillant dans le même rayon (exemple les directeurs du rayon jouet dans les deux succursales).

$$R_{123}' = R_{12}' \text{ [rayon = rayon] } R_3'$$

R_{123}' associe les directeurs de Toulouse, Marseille et Lille travaillant dans le même rayon.

Nous rappelons que la structure de R'_{123} est la suivante :

R'_{123} [dir T, dir M, dir L, rayon, sal T, sal M, sal L]

$R' = (R'_{123} \mid \text{sal T} > \text{sal M et sal T} > \text{sal L})$ [dir T,rayon]

R' : contient le nom des directeurs de Toulouse (ainsi que leur rayon) qui gagnent plus que leurs homologues de Lille et Marseille.

R = (emp T [rayon = rayon] R') [nom,sal]

R : contient le nom des employés (ainsi que leur salaire) satisfaisant la requête.

Nous voyons donc que le traitement de cette requête nécessite la création d'un certain nombre de relations dérivées. Les problèmes majeurs qui se posent sont les suivants :

- les relations R_1, R_2, R_3, R'_{123}, R représentent des joins dont la principale caractéristique est de pouvoir grossir comme un produit cartésien.
Doit-on effectuer ces joins sur le site où se trouvent les deux relations de base (quand c'est possible : cas de R_1, R_2, R_3) ou doit-on envoyer les deux relations de base sans les avoir traitées sur le site où l'on souhaite avoir le join et effectuer cette opération sur place.

- sur quels sites doit-on envoyer les relations intermédiaires ou, le cas échéant les relations de base sachant que le résultat doit être présent à Paris, mais qu'il vaut mieux effectuer une restriction ou une projection sur place et envoyer le résultat sur le site de requête que d'effectuer ces opérations après l'envoi.

Il est clair par exemple, que la relation R'_{123} ne devra jamais être envoyée à Paris, et que si elle y est présente, elle devra y avoir été construite. En effet, cette relation va être considérablement réduite par la sélection et la projection qui vont la transformer en R'.

- Nous pouvons remarquer aussi que, suivant le site où s'effectue la dernière opération de la requête, il peut être intéressant de remplacer dans celle-ci emp T par R_1 si :

1) R_1 n'est pas beaucoup plus grande que emp T
2) R_1 est présente sur le site et emp T doit être transférée de Toulouse.

Cet exemple montre bien que de nombreux problèmes d'optimisation apparaissent si l'on veut minimiser le taux d'occupation des canaux de transfert. Le choix nécessaire des sites de travail et des échanges entre ces sites apparaît comme très critique et une bonne évaluation de la taille des relations dérivées (R_1,R_2,R_3,R'_{12},R'_{123},R') est, dans ce genre de situation, absolument vitale. Cette évaluation a aussi son rôle à jouer lors du choix de l'ordre dans lequel doivent être effectuées certaines opérations. Par exemple, dans la requête nous construisons R'_{12} à partir de R'_1 et de R'_2 puis R'_{123} à partir de R'_{12} et de R'_3, nous aurions pu aussi bien construire R'_{13} (resp. R'_{23}) à partir de R'_1 et R'_3 (resp. R'_2 et R'_3) puis R'_{123} à partir de R'_{13} (resp. R'_{23}) et R'_2 (resp. R'_1). Là encore, une évaluation de la taille de R'_{12}, R'_{13}, R'_{23} est très utile.

Nous nous sommes donc intéressés à la façon d'évaluer ces tailles. Le problème est le suivant : étant donné un certain nombre de relations, une expression relationnelle indiquant comment construire une relation dérivée à partir de ces relations de base, comment évaluer la taille de cette relation dérivée ?

La première étape consiste à établir un modèle mathématique précis caractérisant les propriétés de la base de données ainsi que celles que doit posséder un système de paramètres sur la base afin de permettre une évaluation des relations dérivées à partir d'une requête. C'est l'objet du chapitre I. Le chapitre II contient la définition de deux classes de bases de données pour lesquelles il existe des systèmes de paramètres permettant une évaluation de la taille des relations résultats. Les résultats de ces évaluations sont donnés pour chaque opération de l'algèbre relationnelle et une relation d'ordre entre les différents systèmes de paramètres possibles est présentée. Le chapitre III montre l'importance et le nombre des liens entre les contraintes d'intégrité que peut posséder une base de données et les systèmes de paramètres précédemment définis. Nous montrons dans le chapitre IV que les propriétés que possèdent les deux classes de bases de données définies dans le chapitre II restent conservées lors de l'évolution des bases de données composant ces classes quand une requête est exécutée. En effet, lors de l'évolution d'une requête, un certain nombre de relations nouvelles sont construites à partir des relations de la base. Nous montrons que la nouvelle base constituée de l'ancienne et de ces relations appartient toujours à la classe de départ. Les chapitres V et VI contiennent respectivement la démonstration de la validité de trois systèmes de paramètres sur la première classe de base de données et la validité d'un quatrième système sur la deuxième classe. Nous montrons qu'à l'intérieur de ces deux classes, ces systèmes se "transmettent" à travers les opérations de l'algèbre relationnelle, ce qui permet l'évaluation de requêtes complexes.

II - Le Probleme de l'Optimisation

Comme nous l'avons montré dans l'introduction, le rôle de l'optimiseur dans un système relationnel est très important. Dans ce paragraphe, nous allons indiquer un certain nombre de possibilités et de solutions pour l'optimisation du traitement des requêtes.

Depuis 1971, de nombreux travaux de recherche sur le problème de l'optimisation ont été faits. Certains ont directement abouti à la réalisation d'optimiseurs [Astr 75a] pour SEQUEL ; [Todd 74, Hall 74a, Hall 74b, Hall 75 et Verh 76b] pour PRTV ; [Ston 76 , Wong 76] pour INGRES ; [Blas 76, Blas 77] pour SYSTEM R. D'autres ont eu pour but de transposer des requêtes exprimées à l'aide du calcul relationnel en expressions équivalentes pouvant être implémentées plus aisément [Codd 71c, Pale 72, Pech 75, Roth 74, Roth 75, Tsic 75, Senk 73]. D'autres sont partis de requêtes exprimées à l'aide de l'algèbre relationnelle qu'ils ont transformées en expressions équivalentes pouvant être traitées en parallèle. [Smit 75, Chan 76].

Pour dégager les principales directions, nous allons reprendre et compléter une classification des différentes approches déjà étudiées qui est due à Kim [Kim 79].

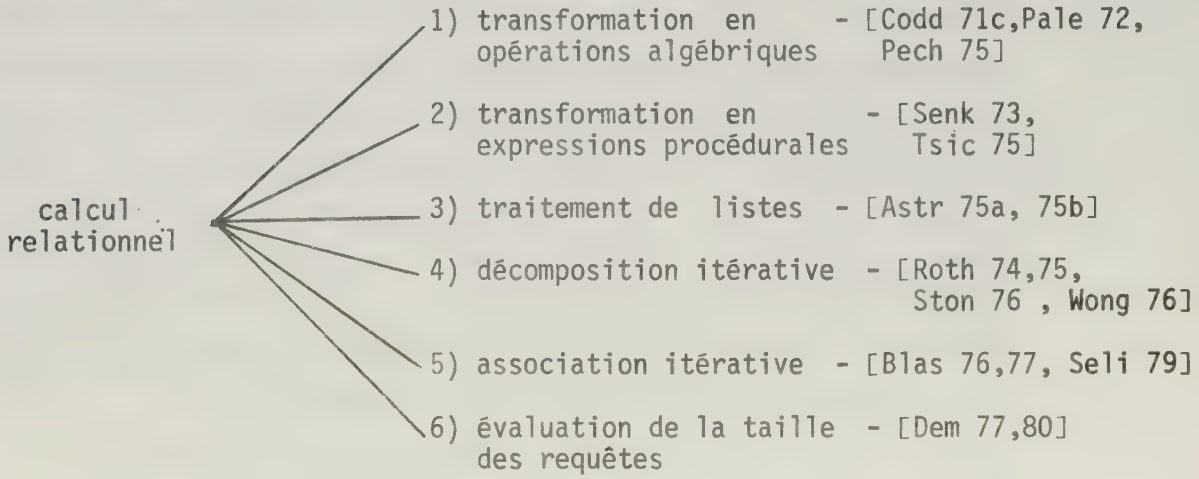
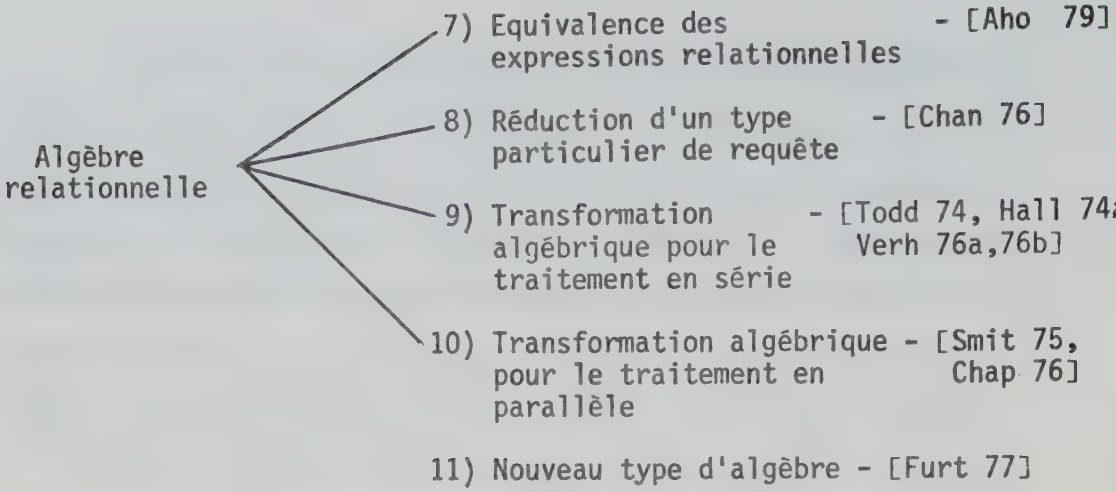


Figure II.1



Nous allons utiliser la Figure II-1 comme plan pour exposer rapidement les différents travaux sur l'optimisation. Codd [Codd 71c] a développé un algorithme transformant une expression du calcul relationnel en suite d'opérateurs algébriques. Palermo [Pale 72] a amélioré l'algorithme et a implémenté les opérateurs de l'algèbre en APL. Pecherer [Pech 75] s'est intéressé à des critères d'efficacité lors de l'implémentation des opérateurs de l'algèbre. Senko [Senk 73] et Tsichritzis [Tsic 75] ont donné des techniques pour transformer une expression non procédurale en suite d'instructions, exprimées dans un langage procédural. Le traitement de listes a été développé par Astrahan et Chamberlin [Astr 75a] dans le cadre du prototype Sequel afin de minimiser le nombre des appels de l'interpréteur Sequel à XRM (Mémoire secondaire). Rothnie [Roth 74,75] donne des techniques permettant de décomposer une expression relationnelle impliquant deux relations en expressions unaires. Wong et Youssefi [Wong 76, Ston 76] ont généralisé ces techniques et les ont incorporées dans INGRES. Blasgen et Eswaran [Blas 76,77] ont développé des techniques pour évaluer des "joins" de relations. Certaines de ces techniques ont été incorporées dans l'optimiseur de SYSTEM R. Demolombe [Dem 80] a étudié les possibilités d'évaluer la taille des relations dérivées de requêtes exprimées par le calcul relationnel en fonction de paramètres statistiques fournis par le système de gestion de la base de données.

Des travaux plus théoriques pouvant déboucher sur des techniques d'optimisation ont été effectués par Aho, Sagiv et Ullman [Aho 78]. Ceux-ci ont étudié l'équivalence des expressions relationnelles contenant les opérateurs de join, projection et sélection. Chandra et Merlin [Chan 76] s'intéressent à une classe particulière de requêtes ("Conjunctive queries") pour laquelle ils donnent des algorithmes permettant la construction de requêtes minimales équivalentes. Hall et Todd [Todd 74, Hall 74a,75] ont étudié des moyens permettant de transformer une requête exprimée dans un langage du type algébrique en une expression équivalente pouvant être évaluée plus efficacement. Hall a aussi développé des techniques pour supprimer des sous-expressions redondantes d'une requête [Hall 74b,75]. Chang et Smith [Smit 75, Chap 76] transforment une requête exprimée en langage algébrique en une suite équivalente d'opérateurs pouvant être traitée en parallèle. Pour ce faire, ils étudient notamment la distributivité des différents opérateurs de l'algèbre entre eux. Furtado et Kerchberg [Furt77] ont une approche résolument différente puisqu'ils définissent un nouveau type d'algèbre (opérant sur des "relations partitionnées") possédant une plus grande souplesse pour la spécification des requêtes et permettant un traitement de requête efficace.

Nous n'insistons pas ici sur les techniques 1,2,3,4,5,7,8,9,10,11 (pour plus de détails voir [Kim 79] et les références des auteurs).

Ce résumé des différentes directions de recherche dans le problème de l'optimisation nous montre que l'aspect évaluation de la taille des relations dérivées d'une expression relationnelle a été peu abordé si ce n'est par Demolombe [Dem 77,80]. Nous ne reviendrons pas sur l'approche de ce dernier qui est statistique et se place dans le cadre du calcul relationnel. Notre approche est probabiliste et se place dans le cadre de l'algèbre relationnelle.

Comme nous l'avons montré, évaluer la taille des relations résultats d'une requête offre pour l'optimiseur un intérêt certain. Cependant, comme nous l'avons constaté, cette approche n'a

pas été très étudiée. Il y a plusieurs raisons à cela. Premièrement, il y a un énorme travail d'optimisation à faire au niveau physique qui ne nécessite pas d'avoir cette possibilité d'évaluation (gestion des index, etc...), de plus, transformer une requête "mal posée" est très payant et ne nécessite pas forcément de connaître la taille des relations intermédiaires. En effet, il est clair qu'effectuer certaines sélections ou projections avant un join est souvent rentable et des travaux en ce sens ont été effectués très tôt [Smit 75, Chap 76]. La plupart des grandes lignes possibles d'optimisation ne nécessitant pas une évaluation fine de la taille des relations dérivées ont donc déjà été assez bien explorées. L'apparition des bases de données réparties et le désir d'une optimisation "plus fine" justifie l'intérêt plus neuf pour l'évaluation de la taille des requêtes.

III - Définitions et Notations

Le modèle relationnel introduit par Codd [Codd 70] a été depuis largement étudié [Codd 71,73,74], [Arms 74,77].

Dans ce paragraphe nous présentons les concepts essentiels de ce modèle et nous précisons les notations utilisées. La notion de base est la notion de relation.

Une relation est définie par un descripteur de relation du type $R(A_1:D_1,\ldots,A_n:D_n)$. R est le nom de la relation, les A_i sont les noms d'attributs et les D_i les domaines qui leur sont associés. Quand il n'y a pas de confusion possible, nous ignorerons les domaines et nous noterons la relation par $R(A_1,A_2,\ldots,A_n)$. L'ensemble des noms d'attributs sera noté par une lettre de la fin de l'alphabet (U,V,W,X,Y,Z). C'est à dire : $U = \{A_1,\ldots,A_n\}$. La relation est donc notée R(U). Un n-uplet est une fonction x de $\{A_1,\ldots,A_n\}$ dans $D_1 \times D_2 \times \ldots \times D_n$ telle que $x(A_i) \in D_i$.

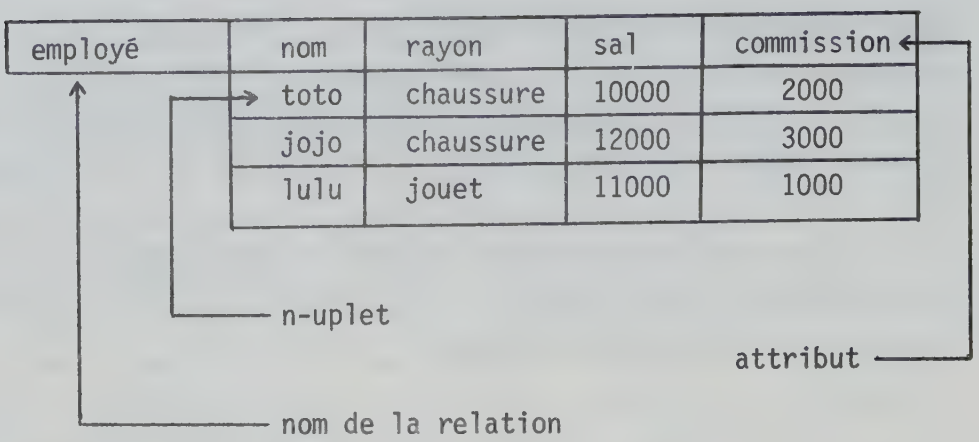
Les valeurs de la relation R sont des ensembles finis de n-uplets, elles sont en général notées en indiçant le nom de la relation : R_t. Ces valeurs sont représentées sous la forme habituelle de table où les lignes correspondent aux n-uplets et les colonnes aux attributs.

La projection d'un n-uplet x sur l'ensemble d'attributs $X \subseteq U$, notée x(X) est la restriction de x à X.

Nous pouvons étendre cette notion au cas des relations. Soit $R_t(U)$ la valeur d'une relation, sa projection sur $X \subseteq U$ notée $R_t(X)$ (ou $R_t(U)[X]$) est définie par : $R_t(X) = \{x(X)/x \in R_t(U)\}$.

Nous clarifions ces notions par l'exemple suivant :

Un magasin tient un fichier de ces employés :



La projection de la relation employée sur l'attribut rayon est :

employé [rayon]	rayon
	chaussure
	jouet

Si $R_t(U)$ et $R'_t(U')$ sont deux valeurs de relations et si x et x' sont deux n-uplets de ces relations, nous appelons concaténation de x et x' et nous notons x!x' le n-uplet x" tel que x"(U) = x(U) et x"(U') = x'(U').

Une base de données est constituée d'une ou plusieurs relations prenant à chaque instant des valeurs spécifiques. Ainsi une base de donnée est décrite par l'ensemble $R = \{(R_1,\ldots,R_p)\}$ de descripteurs de relations. R est le descripteur de la base de données. A ce descripteur nous pouvons associer, à chaque instant t, la valeur de la base de données $R_t = (R_{1t},\ldots,R_{pt})$ où R_{it} est la valeur de la relation R_i.

Dans ce qui suit, nous omettons, lorsqu'il n'y a pas de confusion possible, l'indice t pour désigner les valeurs de relations. L'ensemble des valeurs possibles que ces relations peuvent prendre est le schéma de la base de données noté S. Le schéma est habituellement caractérisé par un ensemble de propriétés qu'une relation doit satisfaire. Ce sont les contraintes d'intégrité, en leur absence, le

schéma est appelé libre.

Nous rappelons ici la définition due à Codd [Codd 72] du type de contrainte que nous utiliserons le plus souvent : les dépendances fonctionnelles définies ainsi : - Soit R(U) le descripteur d'une relation, R_t une valeur de R et $X,Y \subseteq U$. Nous disons que Y dépend fonctionnellement de X dans R_t si et seulement si :

$$\forall x , x' \in R_t,\ x(X) = x'(X) \Rightarrow x(Y) = x'(Y).$$

Cette dépendance est notée $X \underset{R_t}{\rightarrow} Y$ et s'il n'y a pas de confusion possible : $X \rightarrow Y$.

Le schéma caractérisé par un ensemble de dépendances fonctionnelles est l'ensemble des valeurs des relations qui satisfont ces dépendances fonctionnelles.

Nous introduisons une dernière notation [Delobel 78] : Soit $R_t(XY)$ une relation et x un n-uplet de $R_t(X)$; nous notons $R_t(x,Y)$ l'ensemble des n-uplets de $R_t(Y)$ provenant des n-uplets de $R_t(XY)$ dont la valeur sur X est x, soit :

$$R_t(x,Y) = \{y \in R_t(Y) \ / \ x!y \in R_t(XY)\}$$

Nous donnons ici un exemple de base de données et de schéma de cette base :

Soit la base de données "Magasin" constituée des relations suivantes, décrivant un grand magasin :

employé (nom, rayon, salaire, commission)

directeur (nom, rayon)

fournisseur (nom, article, rayon, quantité).

La relation employé donne le nom de chaque employé du magasin, le rayon dans lequel il travaille, son salaire et sa commission.

La relation directeur donne les noms des directeurs de rayon. La relation fournisseur donne le nom des fournisseurs, l'article qu'ils fournissent ainsi que sa quantité et le rayon où il est vendu.

Les domaines de nom, article, rayon sont des alphanumériques de longueur inférieure à 20 caractères. Les domaines de salaire et commission sont les entiers compris entre 1000 et 1000000.

Soient les contraintes d'intégrité suivantes :

- salaire > 2000
- quantité < 10000
- salaire > commission
- article → rayon
- article → quantité
- nom → rayon,salaire, commission

Le schéma de la base de données magasin sera l'ensemble des valeurs possibles des trois relations satisfaisant ces six contraintes d'intégrité.

IV - INTERROGATION DE LA BASE DE DONNÉES ET DÉFINITIONS PRÉLIMINAIRES

Comme nous l'avons montré dans l'exemple étudié dans l'introduction, le résultat d'une requête posée à une base de données relationnelle est une relation (en particulier, il n'y a pas de duplicata). En général, cette relation dérive d'une relation de la base ("donner les noms des employés enregistrés dans la relation EMP T") où de plusieurs relations (voir l'exemple traité en introduction). Le langage de requête doit permettre à l'utilisateur de caractériser la relation qu'il désire construire à partir de la base. Il y a deux types principaux de langages permettant cette caractérisation (Codd 71) :

a) l'utilisateur spécifie la suite d'opérations de l'algèbre relationnelle qui doivent être exécutées afin de construire la relation désirée.

b) Il peut donner les propriétés que doit satisfaire le résultat désiré en termes de calcul relationnel et le système détermine les opérations à effectuer pour obtenir le résultat.

La première approche revient à construire un ensemble en effectuant une suite d'opérations ensemblistes (union, intersection, différence). La deuxième consiste à donner la définition de l'ensemble sous forme de prédicats.

Dans ce qui suit, nous allons nous intéresser à l'évaluation de la taille des relations obtenues par des requêtes exprimées à l'aide de l'Algèbre relationnelle. Nous donnons ici les définitions que nous utiliserons par la suite.

Définition IV-1

Soit R(U) une relation :
Soit E une condition comprenant :

1) comme opérandes des valeurs ou des noms d'attributs

2) les opérateurs arithmétiques de comparaisons :
$<$, $=$, $>$, $\leq$, $\neq$ et $\geq$.

3) les opérateurs logiques $\wedge$, $\vee$ et $\neg$.

Nous appelons sélection de R(U) suivant E (notée R(U)|E) l'ensemble des n-uplets de R(U) vérifiant E. □

Exemple :

Soit R(Employé,Sal. Employé, Directeur, Sal. Directeur) dont une valeur est le tableau ci-après.

Employé	Sal. Employé	Directeur	Sal. Directeur
TOTO	2000	JOJO	4000
JOJO	4000	JOJO	4000
LULU	3000	LILI	5000
LILI	5000	LILI	5000

R(U)|Sal. Employé = Sal. Directeur :

Employé	Sal. Employé	Directeur	Sal. Directeur
JOJO	4000	JOJO	4000
LILI	5000	LILI	5000

R(U)|Sal. Employé = 4000 ou Directeur = LILI

Employé	Sal. Employé	Directeur	Sal. Directeur
JOJO	4000	JOJO	4000
LULU	3000	LILI	5000
LILI	5000	LILI	5000

Définition IV-2

Soient R(U) et R'(U) deux relations.
Soient $X \subseteq U$ et $Y \subseteq V$ deux ensembles d'attributs union compatibles
Soit θ l'un quelconque des comparateurs : $=, \neq, <, \leq, >, \geq$
Le θ join de R sur X et de R' sur Y est noté :

$R[X\theta Y]R'$

C'est la concaténation des lignes x de R et x' de R' telles que leurs composantes sur X et Y vérifient : $x(X) \circ x'(Y)$.

Le cas le plus répandu est l'équi-join (θ est l'égalité) que nous appelons join quand il n'y a pas d'ambiguité.

Exemple :

R	Etudiant	Cours
	TOTO	Maths
	LULU	Maths
	LILI	Latin
	NANA	Anglais

R'	Cours	Professeur
	Maths	ALBERT
	Latin	GEORGES

$R[Cours = Cours]R' = R''(Etudiant, Cours, Professeur)$

R"	Etudiant	Cours	Professeur
	TOTO	Maths	ALBERT
	LULU	Maths	ALBERT
	LILI	Latin	GEORGES

Définition IV-3

Soient R(U) et R'(V) deux relations.
L'union, notée $\cup$, l'intersection, notée $\cap$, et la différence notée - de R(U) et de R'(V) sont définies de la même façon que pour les ensembles

(les éléments ici sont les n-uplets) U et V doivent donc porter sur des domaines de même type. Quand il n'y a pas d'ambiguité nous désignerons par la même lettre les attributs des deux relations. □

Exemple :

R(Employé,Salaire), R'(Directeur, Salaire)

R ∪ R' = R"(Personne, Salaire)

La relation R" représente l'ensemble du personnel de l'entreprise.

Les opérations définies ci-dessus ainsi que l'opération de projection (notée R(U)[X]) définie dans le paragraphe I, forment l'algèbre relationnelle que nous notons $\mathcal{A}$. Pour simplifier certaines démonstrations dans les chapitres suivants, nous allons définir une autre opération qui est un cas particulier de join mais que nous utiliserons telle quelle.

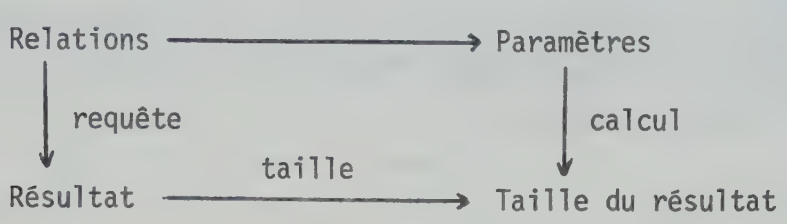
Si R(U) et R(V) sont deux relations, le produit cartésien de R(U) et R(V) (noté R(U) × R(V)) est défini de la même façon que pour les ensembles.

A un schéma donné, nous associons par un des opérateurs de $\mathcal{A}$ la valeur de relation obtenue par application de cet opérateur sur une ou deux quelconques valeurs de relations du schéma.

V - Le Modèle

Nous avons vu dans l'introduction que le problème que nous nous posons est d'évaluer la taille des relations construites lors du traitement d'une requête afin d'optimiser celui-ci.

Nous illustrons par le diagramme suivant la démarche que nous prenons.



Nous voulons estimer la taille du résultat d'une requête avant de l'avoir exécutée. Pour cela il est nécessaire d'avoir un certain nombre de paramètres décrivant les relations de la base de données.

Ces paramètres pourront être par exemple, les espérances des tailles des relations de la base et les degrés de sélectivité des différents attributs de la base. Ces paramètres nous donneront une estimation probabiliste de la taille du résultat d'une requête (espérance ou distribution de taille).

Afin de suivre un approche rigoureuse du problème de l'optimisation, nous devons donc nous donner un modèle mathématique sain et précis afin de décrire la nature probabiliste des relations. Le plus important dans un tel modèle est de définir proprement ce que sont les événements élémentaires.

A/ Schéma probabilisé

Nous suggérons comme étant un choix fécond de considérer comme événements élémentaires les occurences des valeurs de la base de données.

Les domaines dans lesquels les relations de la base de données prennent leurs valeurs, sont dénombrables. Les valeurs des relations sont finies. Le schéma d'une base de données est donc un ensemble dénombrable de valeurs distinctes de la base de données. Nous pouvons donc toujours probabiliser les schémas de la façon suivante :

Définition V-1

Soit S un schéma d'une base de données.

Nous appelons schéma probabilisé le couple (S,p) où p est une mesure de probabilité, tel que :

- $p : S \rightarrow [o,1]$, à R_t associe $p[R = R_t]$. □

A partir d'un tel schéma probabilisé, toutes les autres probabilités peuvent être estimées.

Exemple :

Soit S(Etudiant,Cours)

Soit E l'événement suivant : le nombre de cours suivis par TOTO est de 5.

La probabilité de cet événement est :

$$p(E) = \sum_{R_t \text{ tel que E est vraie dans } R_t} p\,(R = R_t)$$

Il est clair que l'on ne peut définir la mesure de probabilité p dans la pratique. Cette notion formelle nous sert pour définir notre modèle et étudier les différents ensembles possibles de paramètres sur la base de données permettant d'évaluer les tailles des relations dérivées construites à partir des requêtes.

Nous pouvons maintenant définir formellement la notion de système de paramètres.

Nous définissons la notion de système de paramètres comme suit P est une application de C dans l'espace des paramètres où C est un ensemble de schémas probabilisés. Donc, à chaque schéma probabilisé nous associons une valeur des paramètres.

Intuitivement, un tel système définit un ensemble de caractéristiques de la base, pouvant être fournies par le système de gestion de la base de données et mis à jour périodiquement, qui aura pour but de permettre de calculer la taille des relations de la base et surtout d'évaluer la taille de n'importe quelle relation dérivée de ces relations de base par le traitement d'une requête. C'est cette dernière idée qui sera formalisée plus loin dans la définition de "validité" d'un système de paramètres. Nous allons illustrer cette définition par un exemple.

Exemple :

Soit le schéma S(Etudiant,Cours,Professeur)

Les contraintes d'intégrité sont :

Etudiant → Cours

Cours → Professeur

Nous définissons le système de paramètres suivant :

$P : C \rightarrow (\text{réel})^3$

$P : (S,P) \rightarrow (E|R(\text{Etudiant})|, E|R(\text{Cours})|, E|R(\text{Professor})|)$

C est la classe la plus générale,

P est un système de paramètres qui donne l'espérance de la taille de certaines projections des relations de la base. En fait, ce système permet de déduire l'espérance de la taille des relations de la base. En effet, il suffit de remarquer que :

$$|R(\text{Etudiant,Cours,Professeur})| = |R(\text{Etudiant})|$$

car"Etudiant"détermine uniquement tous les n-uplets de cette relation.

De même :

$$|R(\text{Cours,Professeur})| = |R(\text{Cours})|,$$

$$|R(\text{Etudiant,Cours})| = |R(\text{Etudiant})|$$

et

$$|R(\text{Etudiant,Professeur})| = |R(\text{Etudiant})|$$

Par conséquent, ce système de paramètres permet de déduire l'espérance de la taille de toutes les relations construites à partir de la relation de la base par l'opérateur de projection.

Remarque : Dans les applications étudiées, le schéma probabilisé ne sera pas donné en extension. Seul le système de paramètres sera connu. La connaissance de celui-ci ainsi que de certaines conditions que le schéma probabilisé devra vérifier (conditions suggérées par le comportement effectif de la base) caractérisera la mesure de probabilité associée au schéma qui, dans les modèles étudiés, sera simplifiée (propriétés d'équiprobabilité, d'indépendance, etc...).

B/ Validité d'un système de paramètres pour l'algèbre $\mathcal{A}$

Dans ce paragraphe, nous allons donner un certain nombre de définitions et de résultats qui seront utilisés par la suite pour l'étude des différents systèmes de paramètres.

Nous avons indiqué dans le paragraphe II qu'il était nécessaire pour pouvoir effectivement évaluer la taille des relations résultats de requêtes complexes de pouvoir "transmettre" le système de paramètres à travers les opérations effectuées sur la base lors de la réalisation de la requête. Dans ce qui suit, nous allons définir formellement cette notion, cependant, dans le cas général il n'est pas possible d'obtenir de tels systèmes de paramètres, aussi allons-nous définir un certain nombre de conditions sur les schémas des bases de données qui permettent ce "transfert".

Nous appelons classe de schémas probabilisés (notée C) l'ensemble des schémas probabilisés vérifiant un ensemble $\mathcal{C}$ de conditions.

Nous illustrons cette notion par un exemple :

Soit $\mathcal{C}$ la condition d'équiprobabilité suivante :

$\forall R_t \in S, \ \forall \ R_t(U) \in R_t$:

$\forall u,u' \in D(U)$ (où $D(U)$ est le domaine fini de U)

$prob(u \in R_t(U)) = prob(u' \in R_t(U))$

La classe C de schémas probabilisés vérifiant $\mathcal{C}$ est l'ensemble des couples (S,p) où p est une mesure de probabilité telle que $\mathcal{C}$ soit vraie.

Nous notons $S + \Psi(S)$ le schéma contenant les valeurs de relations de S et celles obtenues en appliquant Ψ à une ou deux quelconques valeurs de relations de S.

Nous allons maintenant définir la notion de validité d'un système de paramètres.

Définition V-2

Soit Ψ un opérateur de $\mathcal{A}$

Soit C une classe de schémas probabilisés fermée sous Ψ.

Soit $\underline{P}$ un système de paramètres sur C

$\underline{P}$ est valide dans C pour Ψ si et seulement si :

$\forall S \in C$:

1) $\exists$ f telle que

soit $E(|\Psi(S)|) = f(\underline{P}(S))$

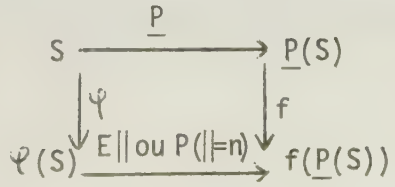
soit $p(|\Psi(S)| = n) = f(\underline{P}(S),n)$

2) $\exists$ g telle que

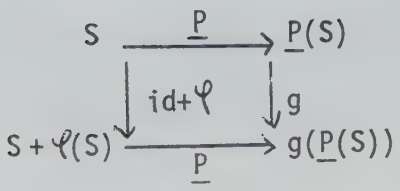
$\underline{P}(S + \Psi(S)) = g(\underline{P}(S))$ □

Cette définition s'illustre par les diagrammes commutatifs suivants :

1) <=>



2) <=>



Intuitivement, nous voyons que (1) exprime le fait que le système de paramètres nous permet de calculer soit l'espérance, soit la distribution de la taille de la relation.

De même (2) signifie qu'à partir des paramètres $\underline{P}(S)$, nous pouvons déduire les paramètres de $\underline{P}(S+\varphi(S))$. En effet, lors de l'évaluation du résultat final d'une requête complexe, il est nécessaire de pouvoir, à chaque opération composant cette requête, calculer les paramètres sur le résultat de cette opération à partir des paramètres de $\underline{P}(S)$ afin d'itérer le processus d'évaluation à chaque étape de la requête. Par exemple, si la requête comprend une sélection suivie d'un join, il est nécessaire d'avoir sur la relation résultat de la sélection les mêmes paramètres que sur la relation de base, ce afin de pouvoir évaluer le résultat du join. Nous savons que la classe de schémas vérifie un certain nombre de propriétés : il faut que ces propriétés restent vraies dans la classe de schémas après application d'une requête. Par exemple, si la requête est une sélection, il faut que le schéma composé de l'ancien schéma et de la nouvelle relation, vérifie toujours ces propriétés si nous voulons pouvoir utiliser les systèmes de paramètres sur ce nouveau schéma. C'est ce que nous entendons par fermeture d'une classe pour une opération.

Le schéma lui-même est un système de paramètres valide pour toutes les opérations de l'algèbre (la classe de schémas est ici l'ensemble infini de tous les schémas possibles).

Nous allons illustrer la définition précédente par un exemple : la classe de schémas considérée est la plus générale puisque l'ensemble des conditions est vide. Soit $\underline{P}$ le système défini de la façon suivante :

$\underline{P}$ associe à tout schéma l'ensemble des espérances des tailles de toutes les projections des relations du schéma (que nous notons $E|R(X)|$)

Soit le schéma S(Etudiant,Cours,Professeur)

$$\underline{P} : S \rightarrow : \underline{P}(S) = (E|R(\text{Etudiant,Cours,Professeur})|, E|R(\text{Etudiant,Cours})|, E|R(\text{Etudiant,Professeur})|, E|R(\text{Cours,Professeur})|, E|R(\text{Etudiant})|, E|R(\text{Professeur})|, E|R(\text{Cours})|)$$

Il est clair que $\underline{P}$ est valide pour l'opérateur de projection. Nous montrons qu'il n'est pas valide pour l'opérateur de sélection. En effet, soit la sélection suivante

$$R'(\text{Etudiant,Cours,Professeur}) = R(\text{Etudiant,Cours,Professeur})|\text{Cours} = \text{Math}$$

$$E|R'| = \begin{cases} 0 \text{ si Math} \notin R(\text{Cours}) \\ \dfrac{E|R(\text{Etudiant,Cours})|}{E|R(\text{Cours})|} \quad \text{si Math} \in R(\text{Cours}) \end{cases}$$

Ici les paramètres ne se "transmettent" pas pour l'opérateur de sélection. En effet, il faudrait connaître :

prob(Math $\in$ R(Cours)).

Or cette probabilité n'est pas dans $\underline{P}(S)$ et ne peut pas en être déduite.

Remarque : Dans la définition VI-3, nous distinguons deux types de systèmes de paramètres, suivant qu'ils déterminent l'espérance ou la distribution de la taille des relations.

C/ Relation d'ordre sur les systèmes de paramètres

Dans ce paragraphe, nous donnons une définition d'une relation d'ordre sur les systèmes de paramètres qui caractérise la finesse relative de l'information qu'ils fournissent sur les schémas sur lesquels ils sont définis.

Définition IV-1

Soit P et P' deux systèmes de paramètres.

Nous disons que P déterminе P' et nous notons $P \geq P'$ si et seulement si :

$\exists f, \forall S$

$P'(S) = f(P(S))$. □

Intuitivement, cette définition indique que les paramètres de P' peuvent se déduire des paramètres de P. Nous allons en donner des exemples :

Soit P le système de paramètres suivant :
$\forall S$,

P associe à S les espérances des tailles de toutes les projections des relations de S.

Soit P_B le système suivant :
$\forall S$,

P_B associe à S les espérances des tailles des projections des relations de S sur les éléments de la base $\mathfrak{B}$ de dépendances fonctionnelles qui leur est associée. Nous démontrons (Chapitre 3, Théorème I-1) que $P_B \geq P$.

Soit P' le système suivant :
∀ S,

P' associe à S les distributions de taille de toutes les projections des relations de S.

Il est clair que P' ≥ P.

Le chapitre III contient de nombreux exemples de systèmes de paramètres équivalent à P ou à P' mais de cardinalité inférieure.

CHAPITRE II

- Résultats -

I - Les classes de schémas probabilisés

II - les systèmes de paramètres (définition)

III - Validité des systèmes de paramètres

Dans ce chapitre nous donnons les résultats sur l'évaluation de la taille des relations construites à partir d'une requête de l'algèbre relationnelle, obtenus grâce aux systèmes de paramètres que nous nous définissons.

Pour cela, nous définissons dans le premier paragraphe deux classes de schémas probabilisés puis dans le paragraphe suivant, nous définissons quatre systèmes de paramètres sur ces classes. Dans le troisième paragraphe, nous présentons les différents résultats obtenus (fermeture des classes pour l'algèbre, validité des systèmes de paramètres pour l'une ou l'autre classe, résultats d'évaluation pour les différentes opérations de l'algèbre relationnelle obtenus grâce à l'un des systèmes).

I - Les classes de schémas probabilisés

A/ Référentiels et conditions

Dans ce sous-paragraphe, nous définissons les différentes conditions qui servent à caractériser l'une ou l'autre de nos classes de schémas probabilisés. Il est clair que dans le cas de schémas probabilisés où la mesure de probabilité est quelconque, nous ne pourrons avoir des résultats d'évaluation intéressants du point de vue de la faisabilité. De plus, dans le monde réel, un certain nombre de propriétés permettent de simplifier d'une manière réaliste les modèles que nous nous proposons de construire.

Nous avons vu dans le premier chapitre, qu'il était nécessaire de pouvoir évaluer des probabilités du type :
prob (5000 ϵ emp(salaire)). En général, la notion de domaine ne nous permet pas d'avoir une évaluation de cette probabilité [Spyr 77]. Pour nos classes de schémas, nous définissons une autre notion qui permet de résoudre ce problème.

Définition I-1

Soit R(U) une relation.

Les référentiels de R(U) (notés $\bar{R}(U)$) sont une collection d'ensembles disjoints satisfaisant les deux axiomes suivants :

A1 : $|\bar{R}(U)| < \infty$

A2 : $\forall\ R_t(U) : R_t(U) \subseteq U\bar{R}(U)$. □

Nous allons, sur un exemple expliciter la notion de référentiel et l'idée de "collection" de référentiels plutôt qu'un référentiel unique.

Soit la relation suivante décrivant les noms, salaires et rayon de travail des employés d'un magasin :

Emp(Nom,Rayon,Salaire)

A ce stade le référentiel de nom est l'ensemble des noms de personnes travaillant dans le magasin.

Celui de rayon est : {jouets, chaussures,vêtements,ménager}

Celui de salaire est l'ensemble des entiers compris entre 1500 et 8000.

Le référentiel de "nom,rayon,salaire" est le produit cartésien des référentiels de chacun des trois attributs.

Considérons l'opération suivante :

R = Emp |Rayon = "chaussures" ∨ Salaire = "2500"

Cette requête peut s'écrire de la façon suivante :

R = (Emp|Rayon = "chaussures") ∪ (Emp|Salaire = "2500")

Le référentiel de (Emp|Rayon = "chaussures") est l'ensemble suivant :

$\overline{Emp}$(Nom) × {chaussures} × $\overline{Emp}$(salaire)

Celui de (Emp|Salaire = "2500") est :

$\overline{Emp}$(Nom) × $\overline{Emp}$(Rayon) × {2500}

Les deux relations formant l'union ne sont pas indépendantes, en effet, l'appartenance de certains n-uplets à l'une n'est pas indépendante de l'appartenance à l'autre. Pour exprimer ces dépendances, nous devons définir les référentiels de R de la façon suivante :

1) $\overline{Emp}(Nom) \times \{chaussures\} \times 2500$

2) $\overline{Emp}(Nom) \times \{chaussures\} \times (\overline{Emp}(Salaire) - \{2500\}$

3) $\overline{Emp}(Nom) \times (\overline{Emp}(Rayon) - \{chaussures\}) \times \{2500\}$

1) contient tous les n-uplets pouvant apparaître dans les deux relations (Emp|Rayon = "chaussures") et (Emp|Salaire = "2500")

2) contient tous les n-uplets ne pouvant apparaître que dans (Emp|Rayon = "Chaussures")

3) contient tous les n-uplets ne pouvant apparaître que dans (Emp|Salaire = "2500")

Durant l'évaluation de la taille des relations dérivées obtenues lors du traitement de la requête, il pourra être nécessaire de créer, à chaque nouvelle relation dérivée une collection de référentiels pour conserver les dépendances inter-relationnelles.

L'intérêt de la notion de référentiel peut se justifier ainsi :

- Nous connaissons généralement l'ensemble des valeurs qui peuvent apparaître dans la base : si nous revenons à l'exemple précédent, nous voyons qu'il est facile de cerner les salaires possibles dans un magasin, les différents rayons peuvent être donnés en extension et nous pouvons facilement savoir quel type de rayon peut exister.

- Les référentiels ne varient pas rapidement : toujours dans l'exemple précédent, la liste des salaires ne changera pas trop souvent.

De même, le magasin ne créera et ne supprimera pas des rayons rapidement par rapport à la vie de la base.

Notation :

Soit (S,p) le schéma probabilisé d'une base de données.
Soit $R(U) \in \mathbf{R}$ et $u \in \cup\bar{R}(U)$
L'événement correspondant à l'ensemble :

$\{ R_t / u \in R_t(U) \}$

est noté $(u \in R(U))$. □

Nota : Nous rappelons que les événements élémentaires sont les occurences des valeurs de la base de données $\mathbf{R}$ (notées R_t).

Nous pouvons maintenant définir les trois types de conditions qui caractérisent les classes de schémas probabilisés considérées par la suite. Nous nous plaçons dans le cas où les relations de la base ont un nombre fini de référentiels.

Définition I-2

Soit (S,p) un schéma probabilisé.
(S,p) satisfait la condition d'équi-probabilité si et seulement si :

$\forall R(U) \in S$:
$\forall X \subseteq U, \forall (x, x') \in \cup\bar{R}(X)^2 : p(x \in R(X)) = p(x' \in R(X))$.

□

Soit l'exemple suivant :

La relation Emp(Nom,Rayon) caractérise le personnel d'un magasin.

Le référentiel de nom est l'ensemble des personnes du magasin. Celui de rayon est {jouets,chaussures,vêtements,alimentation}. La probabilité de l'appartenance de l'un ou l'autre des noms du référentiels de nom ainsi que de l'un ou l'autre des rayons dans la relation emp est la même.

Définition I-3

Nous appelons condition d'indépendance inter-relationnelle la condition suivante :

Soit (S,p) un schéma probabilisé.
Soient R(U) et R'(U') deux relations de ce schéma .
Soient $X \subseteq U$ et $Y \subseteq U'$ deux ensembles d'attributs dont les référentiels sont de même type (attributs union-compatible).

Soient $\bar{R}(X)$ et $\bar{R}'(Y)$ deux référentiels de R(X) et R'(Y)
$\forall \quad (x) \in (\bar{R}(X) \cap \bar{R}'(Y))$:

$(x \in R(X))$ est indépendant de $(x \in R'(Y))$.

Nous dirons que R(X) et R'(Y) sont indépendantes pour $\bar{R}(X)$ et $\bar{R}'(Y)$.

Exemple :

Soient les deux relations suivantes :

Emp(Nom,Rayon), Fournisseur(Nom,Article)

Il est clair que la présence d'une personne portant un nom donné dans Emp(Nom) est indépendante de la présence de son homonyme dans Fournisseur(Nom). Cette condition est liée aux référentiels des relations. En effet, deux relations peuvent être indépendantes pour certains référentiels et pas pour d'autres (voir l'exemple de la définition I.1).

Nous venons de définir l'indépendance de deux relations pour un couple de référentiels non disjoints. Nous allons généraliser cette définition à un ensemble quelconque de relations.

Tout d'abord, nous adoptons la convention suivante : si R(U) et R'(U') sont deux relations telles que, il existe $X \subseteq U$ et $X' \subseteq U'$ tels que X et X' portent sur des domaines union-compatibles alors nous désignons par abus de langage X et X' par la même lettre.

Définition

Soit (S,p) le schéma probabilisé d'une base de données.
Soit $R_x \subseteq R$, un ensemble de relations tel que :

$$X \subseteq \bigcap_{R(U) \in R_x} U$$

(i.e. les R(U) ont des sous ensembles d'attributs dont les domaines sont union-compatibles).

Soit $\bar{R}_x$ un ensemble de référentiels des relations de R_x tel que :

$$\bigcap_{\bar{R}(X) \in \bar{R}_x} \bar{R}(X) \neq \emptyset$$

Si, $\forall x \in \bigcap_{\bar{R}(X) \in \bar{R}_x} \bar{R}(X)$, on a : $p(x \in \bigcap_{R \in \bar{R}_x} R(X)) = \bigcap_{R \in \bar{R}_x} p(x \in R(X))$

alors, nous dirons que les relations de R_x <u>sont indépendantes</u> pour $\bar{R}_x$. □

Nous avons défini ici une notion d'indépendance directement issue de l'indépendance en probabilité ([Rényi 66] p. 51).

Nous allons étendre cette notion à un cadre plus général afin de tenir compte de liens générés entre deux relations par l'opérateur de sélection.

Définition I-5

Soit (S,p) le schéma probabilisé d'une base de données.
Soient R_x et $\bar{R}_x$ définis comme en I-5.
S'il existe R'_x et $\bar{R}'_x$ tels que :

(i) $\bar{R}'_x = \bar{R}_x$

(ii) $\forall x \in \bigcap_{\bar{R}(X) \in \bar{R}_x} \bar{R}(X)$:

$$p(x \in \bigcap_{R(U) \in R_x} R(X)) = p\left(x \in R(X) \,/\, x \in \bigcap_{R'(U) \in R'_x} R'(X)\right) \times \prod_{R(U) \in R_x - \{R_1(U)\}} p(x \in R(X))$$

(où $R_1(U)$ est l'une des relations de R_x).

Nous dirons que les relations de R_x sont indépendantes pour $\bar{R}_x$ modulo (R'_x). □

Pour préciser la troisième condition de manière plus simple, nous posons la définition suivante :

Définition I-6

Soit S un schéma d'une base de données
Soit $R(U) \in \mathbb{R}$
Soit $(u,u') \in (\bar{R}(U))^2$
u et u' sont incompatibles si et seulement si :
$\forall R_t(U)$
$(u \in R_t(U)) \Rightarrow (u' \notin R_t(U))$ et $(u' \in R_t(U)) \Rightarrow (u \notin R_t(U))$. □

Exemple : Soit la relation Emp(Nom,Rayon,Dir)
Les n-uplets :

("toto,jouets,jojo")
et ("toto,chaussure,lili")

ne peuvent coexister dans la même relation car un employé ne travaille que dans un seul rayon. De même les n-uplets :

("toto,jouets,jojo")
("lulu,jouets,nana")

sont incompatibles car un rayon n'a qu'un seul directeur.

Il est clair que la connaissance des contraintes d'intégrité caractérisant le schéma permet de déterminer la compatibilité de deux n-uplets. Pour des raisons de faisabilité, nous ne considérons que des schémas probabilisés dont les contraintes d'intégrité sont des dépendances fonctionnelles (éventuellement des dépendances multi-valuées ou hiérarchiques dans le chapitre III).

Dans ce cadre, nous spécifions dans le lemme suivant le rôle joué par les dépendances fonctionnelles dans la notion de compatibilité.

Lemme I-1

Soit S un schéma caractérisé par sa famille de dépendances fonctionnelles.

Soit R(U) une relation de S.

Soient $(u,u') \in U\bar{R}(U)$.

u et u' sont incompatibles si et seulement si :

$\exists A \subset U, B \subseteq U, A \neq B$ tels que :

$A \rightarrow B \in \mathcal{F}$, $u(A) = u'(A)$ et $u(B) \neq u'(B)$. □

Démonstration :

Celle-ci découle directement de la définition des dépendances fonctionnelles. En effet, s'il existe une dépendance non triviale $A \rightarrow B$ dans $\mathcal{F}$ telle que $u(A) = u'(A)$ et $u(B) \neq u'(B)$ alors toute relation contenant u et u' la brisera et ne pourra donc appartenir au schéma S(U). Inversement si $\forall$ (A, B) tel que $A \rightarrow B \in \mathcal{F}$, $u(A) = u'(A) \Rightarrow u(B) = u'(B)$, la relation composée de u et u' appartient à S, donc u et u' ne sont pas incompatibles.

Nota : La condition d'incompatibilité donnée dans le lemme est, dans tous les cas suffisante. Elle n'est nécessaire que si l'on se restreint, comme nous l'avons dit plus haut, à des schémas de base de données uniquement définis par une famille de dépendances fonctionnelles (i.e. chaque relation composant la base de données n'a pour contraintes d'intégrité que des dépendances fonctionnelles).

B/ Classes de schémas probabilisés

Nous pouvons maintenant définir notre première classe de schémas probabilisés. Pour cette classe, chaque relation de la base devra avoir un nombre fini de référentiels.

Définition I-7

Nous notons CE la classe de schémas probabilisés caractérisée par les propriétés suivantes :

$\forall$ $(S,p) \in CE$,

1) (S,p) vérifie la condition d'équiprobabilité

2) pour tout R_X sous ensemble quelconque de R tel que :

$$X \subseteq \bigcap_{R(U)\in R_X} U$$

Pour tout $\bar{R}_X$ ensemble de référentiels associés aux relations de R_X tel que : $\bigcap_{\bar{R}(X)\in \bar{R}_X} \bar{R}(X) \neq \emptyset$, l'une de deux propriétés suivantes est vérifiée :

i) $\exists\, R'_X$ et $\bar{R}'_X$ définis comme en I.6 tels que les relations de R_X sont indépendantes pour $\bar{R}_X$ modulo (R'_X)

ii) $\exists\, R'_X \subset R_X$ tel que :

$$\left(\bigcap_{R(U)\in R_X} R(X)\right) \cap \left(\bigcap_{\bar{R}(X)\in \bar{R}_X} \bar{R}(X)\right) \supseteq \left(\bigcap_{R(U)\in R'_X} R(X)\right) \cap \left(\bigcap_{\bar{R}(X)\in \bar{R}'_X} \bar{R}(X)\right)$$

3) indépendance intra-relationnelle

$\forall\, R(U) \in R$, $\forall\, X \subseteq U$

$\forall\, (x,x') \in (\cup \bar{R}(X))^2$,

x et x' compatibles $\Longrightarrow$ $(x \in R(X))$ est indépendant de $(x' \in R(X))$.

(R est le descripteur de la base). □

Nota : dans la condition 2, R'_X peut être vide,auquel cas, nous avons indépendance au sens de la définition I.5.

Cette définition caractérise bien une classe de schémas probabilisés. Nous donnerons dans la suite de ce chapitre les trois systèmes de paramètres valides pour cette classe (ainsi que les résultats de l'évaluation des opérations de l'algèbre pour le premier).

La probabilité qu'un n-uplet apparaisse dans une valeur de relation est nécessaire pour l'évaluation comme nous l'avons vu auparavant. Celle-ci peut être l'un des paramètres du système, ou bien peut s'en déduire grâce aux propriétés de la classe de schémas dans le cadre de laquelle est utilisé ce système.

La condition d'équi-probabilité a pour but de permettre l'évaluation de ces probabilités en fonction des paramètres des systèmes que nous étudions (Chapitre IV) dans le cadre de cette classe CE. De même les deux conditions d'indépendance permettent l'évaluation des tailles des résultats d'une opération d'intersection ou de join. Nous allons dans ce qui suit définir une classe plus générale ne possèdant

pas la condition d'équi-probabilité. Le système étudié sur cette classe sera plus complexe car il devra comprendre les probabilités d'appartenance d'un n-uplet à une valeur de relation.

Définition I-8

La classe CI est la classe caractérisée par les propriétés 2 et 3 de la classe CE. □

La collection de référentiels associée à un schéma probabilisé n'est pas uniquement déterminée par la mesure de probabilité.

Aussi adopterons nous la notation suivante : Nous appelons REF la fonction qui à tout (S,p) associe une collection de référentiels.

II - LES SYSTÈMES DE PARAMÈTRES (DÉFINITION)

Le sous paragraphe précédent a défini deux classes de schémas probabilisés. Nous définissons ici quatre systèmes de paramètres dont nous montrerons la validité pour l'une ou l'autre classe de schémas dans les chapitres V et VI. Les deux systèmes de paramètres associés à la classe de schémas notée CE ont pour rôle de permettre l'évaluation de l'espérance de la taille des relations construites à partir d'une requête.

A/ Evaluation de l'espérance de la taille des relations

Définition II-1

Nous notons P le système de paramètres défini de la façon suivante :

Si (S,p) est un schéma probabilisé P associe à S les espéranees des tailles de toutes les projections des relations de S et une valeur de REF. □

Considérons le schéma S constitué des valeurs de la relation

R(Employé,Rayon,Salaire)

$$P(S) = (E|R(\text{Employé})|, E|R(\text{Rayon})|, E|R(\text{Salaire})|, E|R(\text{Employé,Rayon})|,$$
$$E|R(\text{Employé,Salaire})|, E|R(\text{Rayon,Salaire})|,$$
$$E|R(\text{Employé,Rayon,Salaire})|)$$

Ceci est un cas général, nous verrons dans le chapitre III comment restreindre ce système tout en gardant les mêmes possibilités d'évaluation.

Le système P introduit pour une relation portant sur n attributs simples, 2^n-1 paramètres. Il est possible de définir un système de paramètres contenant moins de valeurs.

Définition II-2

Nous notons P_A le système de paramètres défini de la façon suivante :
Si (S,p) est un schéma probabilisé P_A associe à S les espérances des tailles des projections des relations de S sur les attributs simples ainsi que les espérances des tailles des relations et une valeur de REF. □

Si nous considérons le schéma précédent :

$P_A(S) = \{E|R(\text{Employé,Rayon,Salaire})|, E|R(\text{Employé})|,$
$E|R(\text{Rayon})|, E|R(\text{Salaire})|\}.$

Ici nous n'avons besoin que de (n+1) valeurs pour une relation dont l'ensemble d'attributs est composé de n attributs simples.

Il est clair qu'il faudra évaluer l'espérance de la taille des projections des relations du schéma sur des sous-ensembles d'attributs. Nous aurons donc forcément des résultats d'évaluation moins fins que ceux fournis par P. Le système P_A constitue un cas particulier du système P.

B/ Estimation de la distribution de taille des relations

Dans ce paragraphe, nous allons présenter deux systèmes de paramètres permettant d'évaluer les distributions des tailles des relations construites à partir d'opérations de l'algèbre relationnelle. La connaissance de la distribution de taille des relations résultats permet une optimisation plus fine que la connaissance des espérances. En effet, soit $f_R(t)$ la probabilité pour une relation R d'avoir une taille T, si le coût d'une opération donnée sur R est donnée par la fonction C(T) alors nous avons :

$$E(C) = \sum_{T=0}^{\infty} f_R(T)\, C(T).$$

Il est clair que le coût E(C) sera égal à C(E(T)) seulement si C(T) est une fonction linéaire de T. Il est des cas où cela n'est pas vrai, par exemple :

- Accès à la base de données (les fichiers sont généralement implantés hiérarchiquement)
- Tri des n-uplets d'une relation.

Nous allons d'abord présenter un système de paramètres très général puis nous nous restreindrons à un système plus particulier mais plus facilement calculable.

Afin de définir le premier système de paramètres, nous allons introduire un certain nombre de notions caractérisant la structure interne des relations d'un schéma. En effet, il existe des rapports entre le nombre de valeurs d'un ensemble d'attributs donné d'une relation associées à une valeur d'un autre ensemble d'attributs. Les dépendances fonctionnelles en sont un exemple.

Définition II-3

Soit (S(U),p) un schéma probabilisé correspondant aux valeurs de la relation R(U),

Soit $X,Y \subseteq U$

Soit $x \in \upsilon\bar{R}(X)$

Nous notons $f^{P}_{X=X-Y}(n)$ la distribution du nombre d'éléments du domaine de Y associés à l'élément x

Soit $f^{P}_{X=X-Y}(n) = prob(|R(x,Y)| = n)$

$$= \sum_{R_t/n \text{ valeurs distinctes de } Y \text{ associées à } x \text{ dans } R_t} p(R=R_t)$$ □

Nous voyons que ces distributions caractérisent la structure interne des valeurs de relations possibles dans le schéma probabilisé (S(U),p).

Nous allons en donner un exemple :

Soit le schéma S(Etudiant, Cours)
Le domaine de Cours est : {Math,Anglais,Allemand}
Le domaine de Etudiant est : {Toto,Lulu,Lili,Jojo,Nana}

Nous avons le schéma probabilisé suivant :

R_1

Etudiant	Cours
Toto	Math
Lulu	Anglais
Lili	Anglais
Jojo	Allemand

$p(R_1) = 1/8$

R_2

Etudiant	Cours
Toto	Math
Lulu	Math
Lili	Anglais
Jojo	Anglais

$p(R_2) = 1/8$

R_3

Etudiant	Cours
Toto	Math
Lulu	Math
Lili	Math
Jojo	Anglais
Jojo	Allemand

$p(R_3) = 1/4$

R_4

Etudiant	Cours
Toto	Math
Lulu	Math
Lili	Math
Jojo	Math
Jojo	Anglais
Jojo	Allemand

$p(R_4) = 1/2$

ici : $f^p_{\text{Cours = Math - Etudiant}}(1) = 1/8$; $f^p_{\text{Cours = Math - Etudiant}}(3) = 1/4$

$f^p_{\text{Cours = Math - Etudiant}}(2) = 1/8$; $f^p_{\text{Cours = Math - Etudiant}}(4) = 1/2$

$f^p_{\text{Cours = Anglais - Etudiant}}(1) = 7/8$

$f^p_{\text{Cours = Anglais - Etudiant}}(2) = 1/8$

$f^p_{\text{Cours = Allemand - Etudiant}}(0) = 1/8$

$f^p_{\text{Cours = Allemand - Etudiant}}(1) = 7/8$

Remarque : Si (S(U),p) est un schéma probabilisé et X un sous-ensemble quelconque de U, alors $f^{p}_{\emptyset-X}(n)$ a un sens.
Cette distribution est la distribution de taille des projections sur X des relations de S(U) i.e. :

$$f^{p}_{\emptyset-X}(n) = \text{prob}\ \{|R(U)[X]| = n\}\ .$$

Nous pouvons maintenant définir un nouveau système de paramètres.
Nota : Dans le cas général du schéma d'une base de données, les distributions $f_{X=x-Y}(n)$ devraient être indicées par le nom de la relation pour laquelle elles sont définies. Nous ne le faisons pas pour ne pas surcharger les notations.

Définition II-4

Nous notons P' le système de paramètres défini comme suit :

Soit (S,p) un schéma probabilisé

P' associe a (S,p) une valeur de REF et $\{f(i)_{x=x-X}\ ,\ \forall\ X,\ \forall\ x \in U\bar{R}(X)\}$

Ce système de paramètres donne pour tout élément d'un référentiel sa probabilité d'apparaître dans une valeur de relation du schéma. Nous verrons dans le chapitre V qu'il permet de déduire les autres distributions $(f(n)_{X=x-Y})$ qui sont une notation pratique pour exprimer la taille des relations dérivées de certaines opérations de l'Algèbre.

Remarque : La notion de dépendance fonctionnelle est incluse dans celle de distribution $f(n)_{X=x-Y}$.
En effet :

$$(X \rightarrow Y) \iff (f(n)_{X=x-Y} = 0 \quad \forall \quad n > 1, \forall\ x)$$

Nous démontrons dans le chapitre VI que ce nouveau système est valide pour les opérations de l'Algèbre. Il est clair que pour certaines opérations (intersection, union) les calculs sont complexes.

Nous allons maintenant présenter et étudier un autre système de paramètres plus particulier pour lequel les résultats sont plus simples.

Définition II-5

Soit (S(U),p) le schéma probabilisé correspondant aux valeurs de R(U)

Soit X, Y $\subseteq$ U et x $\in$ $\cup\bar{R}(X)$

Soit : $f^{p}_{X=x-Y}(n)$ = prob($|\{y/(xy) \in R(XY)\}| = n$).

Si $f^{p}_{X=x-Y}(n)$ est identiquement distribuée sur l'union des référentiels de R(X), nous la notons : $f^{p}_{X-Y}(n)$. □

Nous nous sommes donc restreints ici à un cas particulier des distributions précédentes. Ce n'est pas une hypothèse d'uniformité sur les distributions, cela signifie simplement que toutes les variables sont identiquement distribuées.

Par exemple, le nombre de cours suivis par un étudiant est identiquement distribué pour tous les étudiants.

Définition II-6

Nous notons P" le système de paramètres défini comme suit

Si (S,p) est un schéma probabilisé, P" associe à S les distributions de taille de toutes les projections des relations de S et une valeur de REF. □

Ici encore, la notion de distribution $f_{X-Y}(n)$ est pratique pour simplifier les notations dans les expressions de taille des relations dérivées lors de la démonstration de la validité de ce système dans le chapitre VI. Nous montrons que pour la classe de schémas CE, $f_{X-Y}(i)$ se construit à partir des paramètres de P".

Nous avons défini quatre systèmes de paramètres différents, Nous allons donner dans le paragraphe suivant les principaux résultats de validité les concernant.

III - VALIDITÉ DES SYSTÈMES DE PARAMÈTRES

Nous rappelons que l'algèbre $\mathcal{A}$ est constituée des opérations de projection, sélection, join, union, intersection, difference.

Théorème III-1

Soit CE la classe de schémas probabilisés définis en I-5.
Soit P le système de paramètres défini en II-1.
P est valide dans CE pour $\mathcal{A}$. □

Démonstration :

La démonstration se fait en deux parties.

Il faut d'abord montrer que CE est fermée pour toutes les opérations de $\mathcal{A}$. C'est l'objet du chapitre IV, puis il faut donner l'expression des paramètres associés à $S + \varphi(S)$, $\forall S \in CE$, $\forall \varphi \in \mathcal{A}$, en fonction de la description de S et de P(S) et prouver que P permet de calculer l'espérance de la taille de la relation construite par φ. Cette deuxième partie est effectuée dans le chapitre V.

Nous allons donner ci-dessous les évaluations des espérances de taille des relations construites par les opérations de $\mathcal{A}$ en fonction des paramètres sur les relations de la base.

Opérations	Paramètres	Espérance de la taille du résultat
projection $R(U)[X]$	$E\|R(X)\|$	$E\|R(X')\|$
sélection $R(U)\|X=x$	$E\|R(U)\|$	$\frac{E\|R(U)\|}{\|U\bar{R}(X)\|}$
sélection $R(U)\|X > x$	$E\|R(U)\|$	$\frac{n\,E\|R(U)\|}{\|U\bar{R}(X)\|}$ où $n = \|\bar{R}(X)\|X > x\|$
sélection $R(U)\|X < x$	$E\|R(U)\|$	$\frac{n\,E\|R(U)\|}{\|U\bar{R}(X)\|}$ où $n = \|\bar{R}(X)\ X < x\|$
sélection $R(U)\|X\neq x$	$E\|R(U)\|$	$X \neq U$ $\frac{E\|R(U)\|}{\|U\bar{R}(X)\|} \times (\|U\bar{R}(X)\|-1)$

Opérations	Paramètres	Espérance de la taille du résultat
sélection $R(U)\|X=Y$	$E\|R(U)\|$	$\frac{E\|R(U)\|}{\|U\bar{R}(XY)\|} \times \|(U\bar{R}(X)) \cap (U\bar{R}(Y)\|$
sélection $R(U)\|X > Y$	$E\|R(U)\|$	$\frac{E\|R(U)\|}{\|U\bar{R}(U)\|} \sum_{y \in U\bar{R}(Y)} n_y$ où $n_y = \|\{x \in \bar{R}(X)/x > y\}\|$
intersection $R'(U) \cap R''(U) = R(U)$	$E\|R'(U)\|$, $E\|R''(U)\|$	1) indépendance inter-relationnelle de $R'(U)$ et $R''(U)$ pour $\overline{R'}(U)$ et $\overline{R''}(U)$ $E\|R(U) \cap (\overline{R'}(U) \cap \overline{R''}(U))\| = \|\overline{R'}(U) \cap \overline{R''}(U)\| \times \frac{E\|R'(U)\| \times E\|R''(U)\|}{\|U\overline{R'}(U)\| \times \|U\overline{R''}(U)\|}$ 2) $R'(U) \cap \overline{R''}(U) \subseteq R''(U) \cap \overline{R'}(U)$ $E\|R(U) \cap (\overline{R'}(U) \cap \overline{R''}(U)\| = \|\overline{R'}(U) \cap \overline{R''}(U)\| \times \frac{E\|R'(U)\|}{\|U\overline{R'}(U)\|}$

Opérations	Paramètres	Espérance de la taille du résultat
Union $R(U) = R'(U) \cap R''(U)$	$E\|R'(U)\|$, $E\|R''(U)\|$	$E\|R(U) \cap (\overline{R'}(U) \cap \overline{R''}(U))\| = E\|R'(U) \cap \bar{R}'(U)\| + E\|R''(U) \cap \bar{R}''(U)\| - E\|R'(U) \cap R''(U) \cap \bar{R}'(U) \cap \bar{R}''(U)\|$
equi-join $R(XYZ) = R'(XY)[X=X]R''(XZ)$	$E\|R'(X)\|$, $E\|R''(X)\|$ $E\|R'(XY)\|$, $E\|R''(XY)\|$	1) indépendance intra-relationnelle dans R'(X) et dans R"(X) et indépendance inter-relationnelle de R'(X) et R"(X) pour $\overline{R'}(X)$ et $\overline{R''}(X)$ $E\|R(XYZ) \cap [(\overline{R'}(X) \cap \overline{R''}(X)) \times U\overline{R'}(Y) \times U\overline{R''}(Z)]\| = \|\overline{R'}(X) \cap \overline{R''}(X)\| \times \frac{E\|R'(XY)\| \times E\|R''(YZ)\|}{\|U\overline{R'}(X)\| \times \|U\overline{R''}(X)\|}$ $E\left\|\left[R(XYZ) \cap [(\overline{R'}(X) \cap \overline{R''}(X)) \times U\overline{R'}(Y) \times U\overline{R''}(Z)]\right][X'YZ]\right\| = \frac{\|U\bar{R}(X'YZ)\|^2}{\|\bar{R}(X'YZ)\|} \left[1 - \left[1 - \frac{E\|R'(XY)\| \times E\|R''(XZ)\| \times \|\bar{R}(X'YZ)\|}{\|U\bar{R}(X'YZ)\|^2 \times \|\overline{R'}(X)\| \times \|\overline{R''}(X)\|}\right]^{\|\bar{R}'(X) \cap \overline{R''}(X)\|}\right]$ où $\bar{R}(X'YZ) = [(\overline{R'}(X) \cap \overline{R''}(X)) \times U\overline{R'}(Y) \times U\overline{R''}(Z)][X'YZ]$ avec $X' \not\supseteq X$ et $X' \neq X$ Remarque : Ce résultat est important dès que l'on a besoin d'une projection de la relation résultat d'un join.

Opérations	Paramètres	Espérance de la taille du résultat
		2) Indépendance intra-relationnelle dans R'(X) et R"(Y) et $R'(X) \cap \overline{R"}(X) \subseteq \overline{R'}(X) \cap R"(X)$.

$$\bar{R}(XYZ) = (\overline{R'}(X) \cap \overline{R"}(X)) \times U\overline{R'}(Y) \times U\overline{R"}(Z)$$

$$E|R(XYZ) \cap \bar{R}(XYZ)| = |\overline{R'}(X) \cap \overline{R"}(X)| \times \frac{E|R'(XY)| \times E|R"(XZ)|}{|U\bar{R}'(X)| \times E|R"(X)|}$$

$$E|[R(XYZ) \cap \bar{R}(XYZ)][X'YZ]| = \frac{|U\bar{R}(X'YZ)|^2}{|\bar{R}(X'YZ)|} \left[1 - \left[1 - \frac{E|R'(XY)| \times E|R"(XZ)| \times |\bar{R}(X'YZ)|}{|U\bar{R}(X'YZ)|^2 \times |U\bar{R}(X)| \times E|R'(X)|}\right]^{|\overline{R'}(X) \cap \overline{R"}(X)|}\right]$$

3ème cas : $U \not\subseteq X$ et $V \subseteq X$ tels que $U \rightarrow V$ dans R'(X) ou R"(X) et indépendance inter-relationnelle entre R'(U,V) et R"(UV).

$$\bar{R}(U) = \overline{R'}(U) \cap \overline{R"}(U) \quad U\bar{R}(V) = (U\overline{R'}(V)) \cap (U\overline{R"}(V))$$

$$E|R(XYZ) \cap \bar{R}(XYZ)| = |\bar{R}(U)| \times |U\bar{R}(V)| \times \frac{E|R'(XY)| \times E|R"(XZ)|}{|U\overline{R'}(UV)| \times |U\overline{R"}(UV)|}$$

Opérations	Paramètres	Espérance de la taille du résultat
		$E\lvert R(XYZ) \cap \bar{R}(XYZ)\rvert[X'YZ] = \frac{\lvert U\bar{R}(X'YZ\rvert^2}{\lvert\bar{R}(X'YZ)\rvert}$ $\left[1 - \left[1 - \frac{\lvert U\bar{R}(V)\rvert \times E\lvert R'(XY)\rvert \times E\lvert R'(XZ)\rvert \times \lvert\bar{R}(X'YZ)\rvert}{\lvert U\bar{R}(X)\rvert \times \lvert U\bar{\bar{R}}''(X)\rvert \times \lvert\bar{R}(X'YZ\rvert^2}\right]^{\lvert\bar{\bar{R}}''(U)\rvert}\right]$
différence $R(U) = R'(U) - R''(U)$	$E\lvert R'(U)\rvert$, $E\lvert R(U) \cap R''(U)\rvert$	$E\lvert R(U)\rvert = E\lvert R'(U)\rvert - E\lvert R'(U) \cap R''(U)\rvert$

Remarque :

Nous ne donnons pas les résultats pour l'opération de join généralisé ($R'(XY)[X\theta X]R''(XZ)$) car celle-ci peut être construite à partir du produit cartésien et des sélections du type $R(U)\restriction X\theta Y$ (cf. Lemme IV.3 du chapitre I).

Nous voyons dans le tableau précédent que ce système de paramètres permet de donner des formules "fermées" pour l'espérance de la taille des relations résultats d'une opération quelconque de l'algèbre relationnelle. La complexité de certaines d'entre elles est surtout due au fait qu'il peut y avoir plusieurs référentiels distincts plutôt qu'un référentiel unique, mais leur nombre n'est pas souvent grand et la plupart des requêtes ne génèrent pas de référentiels multiples ce qui simplifie grandement les résultats.

Théorème III-2

Soit CE la classe de schémas probabilisés défini en I-8.
Soit P_A le système défini en II-2.
P_A est valide dans CE pour $\mathcal{A}$. □

Le théorème est démontré dans le chapitre V.

Théorème III-3

Soit CE la classe de schémas probabilisés définie en I-8.
Soit $\underline{P}''$ le système défini en II-6.
$\underline{P}''$ est valide dans CE pour $\mathcal{A}$. □

Nous savons que CE est fermée pour toutes les opérations de $\mathcal{A}$. La validité de $\underline{P}''$ est démontrée dans le chapitre V.

Nous avons donc un outil permettant d'évaluer les distributions de taille des relations dans le cadre de la classe CE.
Nous généralisons ce résultat dans le cadre de la classe CNE qui n'a pas l'hypothèse d'équi-probabilité d'appartenance des n-uplets à une relation.

Théorème III-4

Soit CNE la classe de schémas probabilisés définie en I-9.
Soit $\underline{P}'$ le système défini en II-4.
$\underline{P}'$ est valide dans CI pour $\mathcal{A}$. □

La démonstration de ce théorème est effectuée dans le chapitre VI.

Nous ne donnons pas de tableau des résultats dans le cas des trois autres systèmes. Ceux-ci figurent dans les chapitres 5 et 6 et ne sont pas en général simples (pour les opérations binaires) principalement pour le système P'.

Tableau récapitulatif des classes et des systèmes :

Classes / Systèmes	CE (condition d'équiprobabilité)	CI
$\underline{P}$: estimation de l'espérance de la taille des requêtes	valide	non valide
P_A : estimation de l'espérance de la taille des requêtes	valide	non valide
$\underline{P}'$: estimation de la distribution de taille des requêtes	*	valide
$\underline{P}''$: estimation de la distribution de taille des requêtes	valide	non valide

De plus nous avons les relations d'ordre suivantes :

$$\underline{P}' > \underline{P}'' > \underline{P} > P_A$$

CHAPITRE III

- Liens avec les contraintes d'intégrité -

I - Relations entre les systèmes de paramètres et les dépendances fonctionnelles

II - Liens avec les dépendances multivaluées et les décompositions du premier ordre

Comme nous l'avons précisé dans l'introduction, une base de données est caractérisée par son schéma qué définit un certain nombre de propriétés que doit posséder la base pour modéliser le monde réel. Ces propriétés sont appelées contraintes d'intégrité. Lors de la définition des trois systèmes présentés dans le chapitre II, nous avons utilisé une contrainte d'intégrité pour définir une des conditions caractérisant les classes de schémas probabilisés dans lesquelles ces systèmes sont valides. Dans ce chapitre, nous allons montrer que la connaissance de cette contrainte d'intégrité (la dépendance fonctionnelle) ainsi que d'autres peut nous permettre de simplifier considérablement l'utilisation de ces systèmes de paramètres ainsi que de minimiser leur cardinalité.

I - RELATIONS ENTRE LES SYSTÈMES DE PARAMÈTRES ET LES DÉPENDANCES FONCTIONNELLES

Nous allons d'abord rappeler quelques résultats importants, dus à Armstrong [Arms 74], sur les dépendances fonctionnelles.

Définition I-1 [Arms 74]

Soit R(U) une relation.

La structure de dépendances de la relation R(U) est la famille :

$$\mathcal{F}_R = \{(X,Y)/X \subseteq U,\ Y \subseteq U,\ X \underset{R}{\rightarrow} Y\} \quad . \qquad \square$$

Armstrong [Arms 74] a isolé quatre axiomes que satisfait toute structure de dépendances d'une relation. (Leur énoncé précis importe peu dans le présent chapitre).

Il est démontré que ce système de quatre axiomes est complet en ce sens que :

- toute structure de dépendances fonctionnelles d'une relation satisfait les axiomes
- pour toute famille de dépendances fonctionnelles (ici, on considère simplement des couples de sous-ensembles de U) satisfaisant ce système de 4 axiomes, il existe une relation R(U) qui admet pour structure de dépendances cette famille.

Nous pouvons maintenant introduire les notions suivantes :

Définition I-2

Soit R(U) une relation.

Soit $\mathcal{F}_R$ sa structure de dépendances.

Soit (X,Y) une dépendance de $\mathcal{F}_R$ (i.e. : $X \xrightarrow{R} Y$)
(X,Y) est dite maximale ssi :

$\forall (X',Y') \in \mathcal{F}_R$ telle que : $X' \subseteq X$ et $Y \subseteq Y'$ alors
$X = X'$ et $Y = Y'$.

L'ensemble des dépendances fonctionnelles maximales de $\mathcal{F}_R$ est appelé famille maximale et notée $\mathcal{M}$. □

Nous pouvons maintenant définir la notion de base d'une structure de dépendances fonctionnelles.

Définition I-3

Soit R(U) une relation.
Soit $\mathcal{F}_R$ sa structure de dépendances fonctionnelles.
Soit $\mathcal{M}$ la famille maximale.
La base de $\mathcal{F}_R$, notée $\mathcal{B}$, est la famille définie comme suit :

$$\mathcal{B} = \{Y \subseteq U / (X,Y) \in \mathcal{M}\} \quad . \qquad \square$$

$\mathcal{B}$ est donc l'ensemble des parties droites des dépendances de $\mathcal{F}$.

Armstrong [Arms 74] a montré que la base est suffisante pour caractériser la structure de dépendances fonctionnelles d'une relation.

Nous allons maintenant pouvoir définir les liens entre ces notions et deux systèmes de paramètres.

Théorème I-1

Soit P le système de paramètres défini en II-1 dans le chapitre II :

P associe à S l'ensemble des espérances des tailles de toutes les projections des relations de S et une valeur de REF.
Soit P_B le système suivant :

P_B associe à S l'ensemble des espérances des projections des relations de S sur les éléments de la base $\mathcal{B}$ et une valeur de REF.

Alors P_B est <u>équivalent</u> à P. □

<u>Démonstration</u> :

P et P_B sont de même type. Il est évident que les paramètres de P_B se déduisent de ceux de P car ($P_B(S) \subseteq P(S)$). Il faut donc montrer que :

$\forall S \in CE$, : $P(S)$ se déduit de $P_B(S)$.

Soit $R(U) \in \mathcal{R}$; Soit $X \subseteq U$ et $\mathcal{B}$ la base de dépendances fonctionnelles de R(U)

1) $X \in \mathcal{B}$, alors $E|R(X)| \in P_B(S)$
2) $X \notin \mathcal{B}$

Ici il faut distinguer deux cas

a) $\exists Y \in \mathcal{B} / X \rightarrow Y \in \mathcal{M}$

Dans ce cas, par définition de $\mathcal{M}$, X est une clef de Y, donc

$E|R(Y)| = E|R(X)|$ et $E|R(Y)| \in P_B(S)$

b) $\forall Y \subseteq U / X \rightarrow Y \notin \mathcal{M}$

donc d'après la définition de $\mathcal{M}$ et les axiomes des dépendances fonctionnelles :
$\forall Z \subseteq U$, $X \rightarrow XZ \notin \mathcal{M}$

Donc il existe $X' \subset X$ et il existe $Y \subseteq U$ tel qu'on ait :

$X' \rightarrow Y \in \mathcal{M}$ et $X \subseteq Y$

donc d'après les axiomes d'Armstrong : $X \rightarrow Y$ et $E|R(X)| = E|R(Y)|$

Comme $Y \in \mathcal{B}$, $E|R(Y)| \in P_B(S)$.

C.Q.F.D.

Remarque :

Ce système P_B est minimal en ce sens que $\forall\ Y \in \mathcal{B}$, $E|R(Y)|$ ne peut être déduite par définition de $\mathcal{B}$ des $E|R(Y')|$ $\forall\ Y' \in \mathcal{B}$ et $Y' \neq Y$.

Nous allons maintenant établir un résultat similaire pour un autre système de paramètres valide.

Corollaire I-1

Soit P" le système de paramètres défini en II-5 dans le chapitre II :

$\forall\ S \in CE$

P" associe à S l'ensemble des distributions de taille de toutes les projections des relations de S et une valeur de REF.

Soit $\underline{P}''_B$ le système défini comme suit :
$\forall\ S \in CE$:

$\underline{P}''_B$ associe à S l'ensemble des distributions de taille des projections des relations de S sur les éléments de la base $\mathcal{B}$ et une valeur de REF.
Alors P''_B est équivalent à P". □

La démonstration de ce résultat est similaire à la démonstration précédente.

II - LIENS AVEC LES DÉPENDANCES MULTIVALUÉES ET LES DÉCOMPOSITIONS DU PREMIER ORDRE

Nous allons d'abord rappeler brièvement les définitions des dépendances multivaluées. [Beeri 77].

Définition II-1

Soit R(U) une relation.

Soit $X, Y \subseteq U$ et $Z = U - XY$.

Nous dirons que Y est en dépendance multivaluée sur X dans R(U) et nous noterons $X \twoheadrightarrow Y$ si et seulement si pour toute valeur $xz \in R(XZ)$ nous avons :

$$R[xz,Y] = R[x,Y]. \qquad \square$$

De cette définition, nous pouvons déduire :

$$E|R(x,Y)| = E|R(xz,Y)|$$

et $\forall\, n : f(n)_{X-Y} = f(n)_{XZ-Y}$.

Donc la connaissance de ces dépendances multivaluées nous permet de simplifier ou d'éviter certains calculs sur les systèmes de paramètres, en particulier lors de l'évaluation de la taille des sélections.

De plus, l'un des axiomes des dépendances multivaluées nous permet d'aller plus avant [Beeri 77].

En effet, si nous avons $X \twoheadrightarrow Y$ dans R(U) alors $\forall\, W \subseteq Z$: $XZ \twoheadrightarrow YW$.

Si on pose $W = \emptyset$, nous avons : $XZ \twoheadrightarrow Y \quad \forall\, Z \subset U-XY$ d'où $f(n)_{XZ-Y} = f(n)_{X-Y}$.

Nous allons maintenant voir les liens avec les décompositions du premier ordre.

Définition II-2 [Delo 78]

Soit R(U) une relation.

Soit $XY_1 \ldots Y_n W = U$.

Nous dirons que $X \rightarrow Y_1 | Y_2 | \ldots | Y_n$ est une décomposition du premier ordre dans R(U) si et seulement si :

(*) $\forall\, x \in R(X) : R(x, Y_1 \ldots Y_n) = R(x, Y_1) \times \ldots \times R(x, Y_n)$. □

De cette définition nous pouvons déduire :

$$\forall\, j \neq i : f_{X-Y_i}(n) = f_{XY_j - Y_i}(n) .$$

De plus nous avons aussi :

$$\forall\, j \neq i, \ell : f_{X-Y_j Y_i}(n) = f_{XY_\ell - Y_i Y_j}(n)$$

et $\forall\, V \subset U - XY_i W$

$$f_{XV-Y_i}(n) = f_{X-Y_i}(n)$$

Ces deux derniers résultats sont des applications de deux axiomes donnés par Delobel [Delo 78]

1) $X \rightarrow Y_1 | Y_2 \ldots | Y_n \Rightarrow X \rightarrow Y_1 Y_2 | \ldots | Y_n$

2) $(X \rightarrow Y_1 | Y_2 \ldots | Y_n)$ et $(Y_1' \subseteq Y_1) \Rightarrow XY_1' \rightarrow Y_1 - Y_1' | Y_2 \ldots | Y_n$

De plus, la relation (*) nous donne une forte relation d'indépendance. En effet soit $(xy_1 \ldots y_n)$ une valeur possible de $R(XY_1 \ldots Y_n)$ alors :

$$P(xy_1 \ldots y_n \in R(XY_1 \ldots Y_n)) = p(xy_1 \in R(XY_1)) \times \ldots \times p(xy_n \in R(XY_n))$$

Nous voyons donc que la connaissance des contraintes d'intégrité de la base a de nombreuses implications au niveau des systèmes de paramètres. En effet, comme nous l'avons montré dans le I, ces contraintes permettent de minimiser la cardinalité des systèmes de paramètres, de plus, elles permettent de simplifier les calculs sur les paramètres lors de l'évaluation de la taille d'une relation créée par une opération de l'algèbre, en ce sens qu'elles créent des liens entre les paramètres d'un système qui n'existent pas sinon.

CHAPITRE IV

- Fermeture des classes -

I - Fermeture de la classe CE pour les opérations de l'algèbre relationnelle

II - Fermeture de la classe CI pour les opérations de l'algèbre relationnelle

Nous avons vu dans le chapitre I que la validité d'un système donné n'a de sens pour une ou plusieurs opérations de l'Algèbre relationnelle que dans une classe de schémas probabilisés fermée pour l'ensemble d'opérations considéré.

En effet, l'ensemble des propriétés qui caractérisent la classe et qui autorisent les calculs sur les paramètres des schémas la composant, ne doit pas être modifié lors de la création d'une nouvelle relation par une opération. Sinon, nous ne pourrions réitérer ces calculs lors des évolutions successives des relations intermédiaires dans une requête complexe.

Il est donc nécessaire avant de montrer que les systèmes de paramètres sont valides pour une classe donnée de montrer que celle-ci est fermée pour les opérations de l'algèbre. C'est l'objet de ce chapitre qui se décompose naturellement en deux parties. La première étudie la fermeture de la classe CE pour les opérations de l'algèbre $\mathcal{A}$, et la seconde étudie la fermeture de la classe CI.

I - FERMETURE DE LA CLASSE CE POUR LES OPÉRATIONS DE L'ALGÈBRE RELATIONNELLE

Nous allons ici donner un certain nombre de propriétés des opérations de $\mathcal{A}$ qui simplifierons les démonstrations de la fermeture des classes de schémas probabilisés ainsi que les démonstrations de validité des systèmes étudiés dans les chapitres V et VI.

Lemme I-1

Soit C une classe de schémas probabilisés.
Soit P un système de paramètres.
Si P est valide dans C pour les opérateurs d'union, d'intersection et de sélection élémentaire, alors P est valide dans C pour l'opérateur de sélection. □

Démonstration :

Les sélections dont l'expression est conjonctive se ramènent à des intersections de relations, celles dont l'expression est disjonctive se ramènent à des unions. Il suffit donc que P soit valide pour ces deux opérations ainsi que pour l'opération de sélection élémentaire.

Lemme I-2

Soit C une classe de schémas probabilisés.
Soit P un système de paramètres.
Si P est valide dans C pour l'opérateur de sélection élémentaire sur l'égalité (i.e. R(U)|X=x), pour les opérateurs d'union et d'intersection, alors P est valide pour les autres types de sélections élémentaires portant sur des attributs dont les domaines associés sont finis. □

Démonstration :

Soit R(U) une relation d'un schéma de la classe.
Soient X, Y $\subseteq$ U tels que :

- X, Y sont associés à des domaines de même type.
- les domaines de X et de Y sont ordonnés et finis.

Nous notons D(X) (resp. D(Y)) le domaine de R(X) (resp. R(Y)).

- Cas de $R(U)|X > x$ où $x \in D(X)$

$$R(U)|X > x = \bigcup_{x' \in D(X)|x' > x} (R(U)|X=x')$$

Il est clair que $\forall\, x', x'' \in D(x)$

$$(R(U)|X=x') \cap (R(U)|X=x'') = \emptyset$$

donc

$$|R(U)|X > x| = \sum_{x' \in D(X)|x' > x} |R(U)|X=x'|$$

Ceci nous permettra d'évaluer facilement l'espérance de la taille d'une telle sélection dans le deuxième chapitre.

- Cas de $R(U)|X < x$, où $x \in D(x)$

Ce cas se traite de façon identique au cas précédent.

- Cas de $R(U)|X \neq x$, où $x \in D(x)$

$$R(U)|X \neq x = \bigcup_{x' \in D(x) - \{x\}} R(U)|X=x'$$

$$|R(U)|X \neq x| = \sum_{x' \in D(x) - \{x\}} |R(U)|X=x'|$$

De même ici dans la classe de schémas particulière présentée dans le chapitre II, ce résultat nous permettra d'évaluer facilement l'espérance de la taille d'une telle sélection.

- Cas de $R(U)|X=Y$, $XY \subseteq U$

$$R(U)|X=Y = \bigcup_{a \in D(X) \cap D(Y)} (R(U)|XY = a!a)$$

soit :

$$|R(U)|X=Y| = \sum_{a \in D(X) \cap D(Y)} |(R(U)|XY = a!a)|$$

- Cas de $R(U)|X > Y$, $XY \subseteq U$

$$R(U)|X > Y = \bigcup_{y \in D(Y)} \left(\bigcup_{x \in D(X)/x>y} R(U)|XY = xy) \right)$$

$$|R(U)|X > Y| = \sum_{y \in D(Y)} \sum_{\substack{x \in D(X) \\ x > y}} |R(U)|XY = xy|$$

Remarque : Ces résultats permettront de prouver la validité d'un système de paramètres pour une classe de schémas. Il est clair que dans le cas général, les calculs d'évaluation qu'ils supposent ne sont pas réalistes. Cependant, dans la classe de schémas appelée CE, ces résultats aboutissent à des calculs simples qui donnent l'espérance de la taille de ces différentes sélections.

Lemme I-3

Le join généralisé est équivalent à un produit cartésien suivi d'une sélection. □

Démonstration :

La démonstration est un résultat de Codd [Codd 71].

Soit R(U) et R'(V) deux relations et soit le θ-join $R[X\theta Y]R'$ où $X \subseteq U$ et $Y \subseteq V$ alors $R[X\theta Y]R' = (R(U).R'(V))[X\theta Y]$.

Ce résultat nous permettra de démontrer facilement la fermeture des classes de schémas probabilisés pour l'opération de join-généralisé.

Nous rappelons la définition de la classe CE.

Définition I-1

Nous notons CE la classe de schémas probabilisés caractérisée par les propriétés suivantes :

$\forall (S,p) \in CE$:

1) (S,p) vérifie la condition d'équiprobabilité

2) pour tout R_X, sous ensemble quelconque de R tel que :

$$X \subseteq \bigcap_{R(U)\in R_X} U$$

pour tout $\bar{R}_X$ ensemble de référentiels associés aux relations de R_X tel que :

$$\bigcap_{\bar{R}(X)\in \bar{R}_X} \bar{R}(X) \neq \emptyset$$

l'une des deux propriétés suivantes est vérifiée :

i) $\exists R'_X$ et $\bar{R}_X$ définis comme en I.6 (Chapitre II) tels que les relations de R_X sont indépendantes pour $\bar{R}_X$ modulo (R'_X)

ii) $\exists R'_X \subset R_X$ tel que :

$$\left(\bigcap_{R(U)\in R'_X} R(X)\right) \cap \left(\bigcap_{\bar{R}(U)\in \bar{R}'_X} \bar{R}(X)\right) \subseteq \left(\bigcap_{R(U)\in R_X} R(X)\right) \cap \left(\bigcap_{\bar{R}(U)\in \bar{R}_X} \bar{R}(X)\right)$$

3) indépendance intra-relationnelle

$\forall R(U) \in R$, $\forall X \subseteq U$

$\forall (x,x') \in (\cup \bar{R}(X))^2$: x et x' compatibles $\Longrightarrow$ $(x \in R(X))$ est indépendant de $(x' \in R(X))$. □

Démontrer la fermeture de cette classe, consiste à montrer que la nouvelle relation obtenue par application d'un opérateur de l'algèbre sur une ou deux relations d'un schéma quelconque de la classe vérifie, ainsi que ses référentiels qui sont construits à partir des référentiels des relations opérandes, toujours les trois propriétés de la classe soit les conditions d'équiprobabilité, d'indépendance inter-relationnelle et d'indépendance intra-relationnelle.

Nous allons le montrer pour chaque opération de l'algèbre en commençant par l'opération de sélection. En effet, il est clair d'après les propriétés de la classe CE que celle-ci est fermée pour l'opérateur de projection.

Théorème I-1

La classe CE est fermée pour l'opération de sélection élémentaire sur l'égalité $(R(U)|X=x)$. □

Démonstration :

Soit $S \in CE$, $R \in S$ et $R(U) \in R$.

Soit $R'(U)$ la relation construite ainsi : $R'(U) = R(U)|X=x$ ∀ $\bar{R}(U)$ un référentiel de $R(U)$: $\bar{R}(U)|X=x$ sera un référentiel de $R'(U)$.

Nous montrons que la propriété 2 de la classe CE reste vraie pour la nouvelle base obtenue en rajoutant $R'(U)$ à R.

1er cas : Soit $R_X \subseteq R$, l'ensemble des $R_i(U_i) \in R$ tel que :

$$\bigcap_{i=1}^{n} U_i = X$$

Il est clair que $R(U) \in R_X$.

Soit $\bar{R}_1(X),\ldots,\bar{R}_n(X)$ un ensemble de référentiels des relations de R_X tel que :

$$\bigcap_{i=1}^{n} \bar{R}_i(X) \neq \emptyset$$

Si $x \notin \bigcap_{i=1}^{n} \bar{R}_i(X)$ la démonstration est finie. Nous supposons donc que $x \in \bigcap_{i=1}^{n} \bar{R}_i(X)$.

Si nous rajoutons $R'(X)$ à l'ensemble R_X, il est clair que $\left(\bigcap_{i=1}^{n} \bar{R}_i(X)\right) \cap \bar{R}'(X) = \{x\}$.

Si les relations de R_X sont indépendantes, alors $(x \in R_i(X))$ est une suite d'événements mutuellement indépendants pour i = 1 à n.

Par définition de la sélection :

$\forall x \in \bar{R}'(X),\quad x \in R'(X) \Longleftrightarrow x \in R(X)\ (\bar{R}'(X) = \{x\})$

Donc : $p(x \in R'(X) \wedge x \in \bigsqcap_{i=1}^{n} R_i(X)) =$

$$p(x \in R(X) \wedge x \in \bigsqcap_{i=1}^{n} R_i(X)) = \prod_{i=1}^{n} p(x \in R_i(X))$$

Dans le cas d'indépendance modulo R_X, la démonstration est similaire.

Si il existe $L = \{k_1,\ldots,k_\ell\}$ tel que $|L| < n$ et :

$$(\bigsqcap_{i=1}^{n} R_i(X)) \cap (\bigsqcap_{i=1}^{n} \bar{R}_i(X)) \supseteq (\bigsqcap_{k\in L} R_k(X)) \cap (\bigsqcap_{k\in L} \bar{R}_k(X))$$

Si $R(X) \in \{R_{k_1}(X),\ldots,R_{k_\ell}(X)\}$

Clairement :

$$R'(X) \cap \bar{R}(X) \cap (\bigsqcap_{i=1}^{n} R_i(X)) \cap (\bigsqcap_{i=1}^{n} \bar{R}_i(X)) \supseteq (\bigsqcap_{k\in L} R_k(X)) \cap (\bigsqcap_{k\in L} \bar{R}_k(X))$$

Si $R(X) \notin \{R_{k_1}(X),\ldots,R_{k_\ell}(X)\}$

Alors nous sommes ramené au début de la démonstration en remplaçant R_X par $\{R_{k_1},\ldots,R_{k_\ell},R\}$.

Comme $|R_X| < \infty$, la démonstration aboutit.

<u>2ème cas</u> : Soit $R_Y \subseteq R$ l'ensemble des $R_i(U_i) \in R$ tel que $\bigsqcap_{i=1}^{n} U_i = Y$ avec $Y \cap X = \emptyset$

Nous supposons que $R(U) \in R_Y$ (sinon, il suffit de considérer l'ensemble $R_Y \cup \{R(U)\}$).

Nous devons étudier le comportement de :

$R_Y \cup \{R'(U)\}$

A) Posons $R^n(Y) = \bigcap_{i=1}^{n} R_i(Y)$. Supposons que : $\forall\, y \in \bigcap_{i=1}^{n} \bar{R}_i(Y)$:

$$p(y \in R^n(Y)) = \prod_{i=1}^{n} p(y \in R_i(Y)) \qquad (*)$$

(Le cas où il y a indépendance modulo $R'_Y \neq \emptyset$ se traite de façon similaire).

Nous cherchons : $p(y \in R'(Y) \cap (R^n(Y)))$.

Nous savons que : $(y \in R'(Y)) <==> (xy \in R(XY))$

donc :

$$p(y \in R'(Y) \cap R^n(Y)) = p(xy \in R(XY) \wedge y \in R^n(Y)).$$

D'après la relation (*), nous avons :

$$p(xy \in R(XY) \text{ et } y \in R^n(Y)) = p(xy \in R(XY)) \times p(y \in R^n(Y))$$

soit : $p(y \in R'(Y) \cap R^n(Y)) = p(y \in R'(Y)) \times p(y \in R^n(Y))$.

B) Supposons que :

$\exists L = \{k_1,\ldots,k_\ell\}$ tel que

$R^L(Y) \cap \bar{R}^L(Y) \subseteq R^n(Y) \cup \bar{R}^n(Y)$

(où R^L, $\bar{R}^L$, $\bar{R}^n$ sont définis comme R^n).

Alors le problème se ramène à étudier le comportement de $\{R'(Y), R_{k_1}(Y),\ldots,R_{k_\ell}(Y)\}$.

Soit nous sommes ramené au cas précédent, soit il existe $L' = \{k'_1,\ldots,k'_\ell\}$ tel que $|L'| < |L|$ et $\bar{R}^{L'}(Y) \cap R^{L'}(Y) \subseteq R^L(Y) \cap \bar{R}^L(Y)$

Par itération, nous arrivons aux deux cas suivants :

a) $R(Y) \cap \bar{R}(Y) \cap \bar{R}''(Y) \subseteq R(Y) \cap R''(Y) \cap \bar{R}(Y) \cap \bar{R}''(Y)$

b) $R''(Y) \cap \bar{R}(Y) \cap \bar{R}''(Y) \subseteq R(Y) \cap R''(Y) \cap \bar{R}(Y) \cap \bar{R}''(Y)$

où $R''(Y)$ (resp. $\bar{R}''(Y)$) désigne, non pas forcément, une relation de la base, mais l'intersection d'un certain nombre d'entre-elles.

Dans le cas a), nous avons puisque $R'(Y) \subseteq R(Y)$ et $\bar{R}'(Y) = \bar{R}(Y)$:

$$R'(Y) \cap \bar{R}'(Y) \cap \bar{R}''(Y) \subseteq R'(Y) \cap R''(Y) \cap \bar{R}'(Y) \cap \bar{R}''(Y)$$

Dans le cas b) : $\forall\, y \in \bar{R}'(Y) \cap \bar{R}''(Y)$

$$p(y \in R'(Y) \cap R''(Y)) = p(y \in R'(Y)) \times p(y \in R''(Y)/y \in R'(Y))$$

Par définition de $R'(Y)$:

$$p(y \in R'(Y) \cap R''(Y)) = p(y \in R'(Y)) \times p(y \in R''(Y)/xy \in R(XY))$$

$R'(Y)$ et $R''(Y)$ sont donc indépendants pour $\bar{R}'(Y)$ et $\bar{R}''(Y)$ modulo $(R(Y))$.

La probabilité $p(y \in R''(Y)/xy \in R(XY))$ peut être estimée d'après les propriétés de la classe.

Si nous supposons que :

$$p(xy \in R(XY)/y \in R''(Y)) = p(xy \in R(XY)/y \in R(Y)) :$$

$$p(y \in R''(Y)/xy \in R(XY)) = \frac{p(xy \in R(XY)/y \in R''(Y) \times p(y \in R''(Y))}{p(xy \in R(XY))}$$

$$= \frac{p(y \in R''(Y))}{p(xy \in R(XY))} \times p(xy \in R(XY)/y \in R(Y))$$

$$p(y \in R''(Y)/xy \in R(XY)) = \frac{p(y \in R''(Y))}{p(y \in R(Y))}$$

Le cas où $Y \cap X \neq \emptyset$ se traite de la même façon. Le cas où $X \subseteq Y$ se traite comme le premier cas.

La condition 3 est toujours vérifiée par R'.

Théorème I-2

La classe CE est fermée pour l'opération d'intersection. □

Démonstration :

$\forall (S,p) \in CE$, $\forall R'(U), R''(U) \in S$.

Soit R(U) la relation construite de la façon suivante :

$R(U) = R'(U) \cap R''(U)$.

La collection de référentiels de R(U) est l'ensemble des référentiels $\bar{R}(U)$ obtenus de la façon suivante :

$\forall \bar{R}'(U)$, $\forall \bar{R}''(U)$ deux réferentiels non disjoints de R'(U) et de R"(U)

$\bar{R}(U) = \bar{R}'(U) \cap \bar{R}''(U)$.

Il faut vérifier que la propriété 2 est conservée.

Soit $R_U \subseteq R$ l'ensemble des relations $R_i(U_i)$, $1 \leq i \leq n$ tel que : $\bigcap_{i=1}^{n} U_i = U$.

Il est clair que R'(U) et R"(U) appartiennent à R_U

Par définition de R(U), la propriété (2) est conservée.

En effet, $\forall R'_U \in R_U$, nous avons deux cas :

1) R'(U) et R"(U) $\in R'_U$; dans ce cas, le comportement de $R'_U \cup \{R(U)\}$ est celui de R'_U.

2) R'(U) ou R"(U) $\notin R'_U$; dans ce cas, d'après la propriété (2), il existe $R''_U = R'_U \cup \{R'(U), R''(U)\}$ dont on connait le comportement. Le comportement de $R'_U \cup \{R(U)\}$ est celui de R''_U par définition de R(U).

Théorème I-3

La classe CE est fermée pour les opérateurs d'union et de différence. □

Démonstration :

(Dans ce qui suit on ne représente pas les ensembles d'attributs qui sont communs aux relations étudiées).

$\forall$ (S,p) $\in$ CE

$\forall$ R' et R" $\in$ $\mathbb{R}$

Soit R = R' $\cup$ R"

Nous allons d'abord définir la collection des référentiels de R. Soit $\bar{R}'$ l'union des référentiels de R' et $\bar{R}"$ l'union des référentiels de R".

Soit $\bar{R}'_1$ un référentiel de R'.

1) Soit $\bar{R}'_1 \subseteq \bar{R}' - \bar{R}"$. Dans ce cas, ce référentiel reste inchangé et devient un référentiel de R.

2) Soit $\bar{R}'_1 \subseteq \bar{R}' \cap \bar{R}"$. Dans ce cas, pour chaque référentiel de R", on crée un référentiel de R en considérant son intersection avec $\bar{R}'_1$.

3) Dans le cas où $\bar{R}'_1$ ne vérifie aucune des deux propriétés précédentes, le sous-ensemble de $\bar{R}'_1$ qui est inclus dans $\bar{R}' - \bar{R}"$ devient un référentiel de R et son complémentaire est traité comme dans le cas 2.

Il reste à compléter la collection en rajoutant les référentiels de R" inclus dans $\bar{R}" - \bar{R}'$.

Nous avons bien entendu $\bar{R} = \bar{R}' \cup \bar{R}"$

La conservation de la propriété deux à travers l'union découle des résultats pour l'intersection.

En effet, nous avons partitionné les référentiels de R en référentiels dont les éléments venaient soit de $\bar{R}' - \bar{R}"$, soit de $\bar{R}" - \bar{R}'$ soit de $\bar{R}' \cap \bar{R}"$.

Les référentiels des deux premiers types satisfont par construction la propriété 2 pour ceux de R' $\cap$ R" le résultat a été démontré dans l'étude de l'intersection.

Il reste à vérifier que la propriété trois est toujours vraie.

1er cas : $\forall (x,x') \in (\cup \bar{R}') \cup (\cup \bar{R}")$

$x \in R'$ est indépendant de $x' \in R"$

et

$x \in R"$ est indépendant de $x' \in R"$

Dans ce cas on a aussi l'indépendance dans l'union.

2ème cas : il existe une relation de dépendance entre les attributs d'une des deux relations.

Soit il n'en existe pas dans l'autre et l'union brise la dépendance fonctionnelle, d'où indépendance intra-relationnelle. Soit : dans R'(U) :

$$X,X',Z/Z \subset X \cap X' \text{ et } Z \neq X \text{ et } X' \;/\; x[Z] = x'[Z] \text{ et } Z \rightarrow ZW$$
$$\text{où } W \subset X \cap X' \text{ et } W \neq \emptyset$$

et dans R"(U)

$$Z' \;/\; Z' \subset X \cap X' \text{ et } Z' \neq X \text{ et } X' \;/\; x[Z'] = x'[Z'] \text{ et } Z' \rightarrow Z'W'$$
$$\text{où } W' \subset X \cap X' \text{ et } W' \neq \emptyset$$

alors si $Z' \rightarrow ZW$ (ou $Z \rightarrow Z'W'$), la condition de dépendance reste vraie dans $R' \cup R"$. Sinon les deux dépendances fonctionnelles sont brisées et il y a indépendance entre X et X'.

La validité de CE par la différence se montre de la même façon que pour l'intersection.

Théorème I-4

La classe CE est fermée pour l'opération de sélection. □

Démonstration :

Celle-ci est une conséquence directe des lemmes VI-1 et VI-2 du chapitre I et des théorèmes I-1, I-2, I-3.

En effet,CE est fermée pour l'union, l'intersection et la sélection élémentaire du type R(U)|X=x. Nous sommes donc dans les conditions d'application du lemme VI-1 qui montre que toute sélection peut être considérée comme une requête impliquant ces trois opérateurs.

Théorème I-5

La classe CE est fermée pour l'opération de join généralisé. □

Démonstration :

Soit S ∈ CE,R'(XY) et R"(XZ) deux relations de S
Soit R(XYZ) = R'(XY)[XθX]R"(XZ) .

Le lemme VI-3 du chapitre I nous permet de considérer le join généralisé comme un produit cartésien suivi d'une sélection. Nous savons que CE est fermée pour la sélection. Il reste à voir que les trois propriétés sont conservées par le produit cartésien.

Soit (S,p) ∈ CE, R'(U) et R"(V) ∈ S.
Soit R(UV) = R'(U) ⊗ R"(V).

Pour chaque couple de référentiels de R'(U) et de R"(V) nous créons un référentiel de R(UV) en effectuant le produit cartésien des référentiels de ce couple.

Il est clair que la propriété 1 reste vraie.
∀ X ⊆ U : R(UV)[X] = R'(U)[X] et ∀ X ⊆ V : R(UV)[X] = R"(V)[X]

Les propriétés des projections de R(UV) (tant inter-relationnelle que intra-relationnelle) sont donc bien conservées.

Nous avons donc démontré que la classe CE est fermée pour les opérations de l'algèbre relationnelle puisque pour tout schéma de la classe CE,les nouvelles valeurs de relations créées par les opérations de l'algèbre vérifient toujours les trois propriétés.

II - FERMETURE DE LA CLASSE CI POUR LES OPÉRATIONS DE L'ALGÈBRE RELATIONNELLE

La classe CI a été définie en I.9 dans le chapitre II. Cette classe diffère de la classe CE par le fait qu'elle ne contient pas d'hypothèse d'équi-probabilité.

Théorème II-1

La classe CI est fermée pour les opérations de l'algèbre relationnelle. □

Démonstration :

La classe CI est une généralisation de la classe CE puisque elle vérifie les mêmes conditions moins une.

Les démonstrations du paragraphe précédent pour la fermeture de CE s'appliquent puisqu'elles ne font pas appel à la propriété d'équiprobabilité.

Dans ce chapitre, nous avons donc montré que les deux classes présentées dans le chapitre II sont fermées pour les opérations de l'algèbre relationnelle. Ce résultat est très important, car il nous assure que les relations obtenues à chaque étape de l'exécution d'une requête complexe continuent à vérifier les propriétés du schéma auquel appartiennent les relations dont elles sont dérivées. Aussi lors de l'évaluation de la taille des résultats d'une requête, nous pourrons toujours utiliser ces propriétés pour effectuer les calculs et ce à toutes les étapes constituant la requête.

CHAPITRE V

- Validité des systèmes P, P_A et P" -

I - Validité des systèmes P et P_A dans la classe CE pour les opérations de l'algèbre relationnelle

II - Validité du système P" dans la classe CE pour les opérations de l'algèbre relationnelle

Dans ce chapitre, nous allons étudier la validité des systèmes de paramètres P et P_A (définis dans le chapitre II) dans la classe CE pour les opérations de l'algèbre. La démonstration de la validité est constructive. En effet, pour chaque opération de l'algèbre (projection, sélection, union, intersection, join) nous donnons une formule fermée de l'espérance de la taille du résultat d'une part et de toutes les projections de la relation résultat d'autre part, ce afin de prouver que le système de paramètres P permet non seulement d'évaluer l'espérance de la taille de la relation résultat , mais encore que les paramètres se "transferrent" à travers les opérations. Nous montrons que le système P_A se traite comme un cas particulier du système P.

I - VALIDITÉ DES SYSTÈMES P ET P_A DANS LA CLASSE CE POUR LES OPÉRATIONS DE L'ALGÈBRE RELATIONNELLE

Nous rappelons ici la définition du système de paramètres P.

Définition I-1

Nous notons P le système de paramètres suivant :

Si (S,p) est un schéma probabilisé :

P associe à S l'ensemble des espérances de taille de toutes les projections des relations de S et une valeur de REF.

La définition de la classe CE est donnée en I-8 dans le chapitre II.

Nous rappelons que CE est fermée pour les opérations de l'algèbre.

Théorème I-1

P est valide dans CE pour l'opérateur de projection. □

Démonstration :

$\forall (S,p) \in CE, \forall R_t \in S, \forall R_t(U) \in R_t, \forall X \subseteq U$

$E|R_t(X)| \in P(R_t)$

P est donc valide pour la projection dans CE.

Théorème I-2

P est valide pour l'opération de sélection élémentaire sur l'égalité dans CE. □

Démonstration :

Soit $(S,p) \in CE$, soit $R(XY) \in S$

Soit la sélection $R(XY)|X = x$

Nous allons étudier les divers cas possibles :

1) R(XY) vérifie la condition d'indépendance intra-relationnelle
   $xy \in R(XY)$ est indépendant de $xy' \in R(XY)$
   Dans ce cas, chercher à déterminer $prob(|R(x,Y)| = n)\ \forall n$ revient à chercher à déterminer la probabilité d'avoir n couples (xy) dans R(XY), chacun ayant la même probabilité

de s'y trouver. Ces couples sont choisis parmi : {x} X(U$\bar{R}$(Y)).
Nous nous trouvons donc devant une distribution binomiale

$$\text{prob}(|R(x,Y)| = n) = C^{n}_{|U\bar{R}(Y)|} \left(\frac{E|R(XY)|}{|U\bar{R}(XY)|}\right)^{n} \times \left(1 - \frac{E|R(XY)|}{|U\bar{R}(XY)|}\right)^{|U\bar{R}(Y)| - n}$$

2) $\exists(U,V)$ une partition de (X,Y) telle que $U \rightarrow V$ dans R(XY).

a) $U = X$ et $Y = V$;

Dans ce cas, puisque $X \rightarrow Y$ dans R(XY)

$$E|R(x,Y)| = \text{prob}((xy) \in R(XY)) = \frac{E|R(XY)|}{|U\bar{R}(XY)|} \times |U\bar{R}(Y)| = \frac{E|R(XY)|}{|U\bar{R}(X)|}$$

b) $Y = Y'Y''$ où $Y' = Y \cap V$

alors : $XY'' \rightarrow Y'$ puisque $U \subseteq XY''$

dans ce cas, le nombre de valeurs de Y pour X sera le nombre de valeurs de Y" pour X.

Nous savons que $X \not\rightarrow Y''$ dans R(XY) car alors Y" serait dans V ce qui contredirait l'hypothèse.

Nous sommes donc ramené au premier cas.

Il ne nous reste plus qu'à calculer E|R(XY)|.

Soit :

$$E|R(x,Y)| = \sum_{i=1}^{|\bar{R}(Y)|} C^{n}_{|\bar{R}(Y)|} \left(\frac{E|R(XY)|}{|U\bar{R}(XY)|}\right)^{i} \times \left(1 - \frac{E|R(XY)|}{|U\bar{R}(XY)|}\right)^{|U\bar{R}(Y)| - i}$$

d'où :

$$E|R(x,Y)| = \frac{E|R(XY)| \times |U\bar{R}(Y)|}{|U\bar{R}(XY)|} = \frac{E|R(XY)|}{|U\bar{R}(X)|}$$

Pour terminer la démonstration, il faut montrer que
$\forall\ Z \subseteq XY$: E|R(x,Y) [Z]| se déduit de $P(R_t)$

1er cas : $Z \subseteq X$: E|R(x,Y) [Z]| = prob($x \in R(X)$)

2ème cas : $Z \not\subseteq X$: E|R(x,Y) [Z]| = R(x,Z).

Nous nous ramenons au cas général.

Théorème I-3

P est valide pour CE pour l'opération de sélection élémentaire du type R(U)|X > x, R(U)|X < x, où R(U)|X ≠ x. □

Démonstration :

Soit R(U) une relation d'un schéma de CE ; nous savons (Lemme VI-2 du chapitre I) que :

$$|(R(U)|X > x)| = \sum_{x' \in U\bar{R}(X)|x'>x} |R(U)|X = x'|$$

si $|(U\bar{R}(X)|X > x)| = n_x$ alors, puisque $E|(R(U)|X = x)|$ est indépendante de x :

$$E|(R(U)|X > x)| = n_x \times E|(R(U)|X = x)|$$

$E|(R(U)|X = x)|$ a été calculée dans le théorème précédent.

Le second cas de sélection se traite de façon similaire.

Le troisième cas, $R(U)|X \neq x$ a pour résultat

$$E|(R(U)|X = x)| = |(U\bar{R}(X)|X \neq x)| \times E|R(U)|X = x|$$

$$= (|U\bar{R}(X)| - 1) \times E|R(U)|X = x|$$

Théorème I-4

P est valide dans CE pour les sélections du type $R(U)|X \ominus Y$ où $\ominus \in \{=, >, <\}$. □

Démonstration :

1) $R(U)|X = Y$, $XY \subseteq U$

D'après le Lemme VI-2 du chapitre I :

$$|R(U)|X = Y| = \sum_{a \in (U\bar{R}(X)) \cap (U\bar{R}(Y))} |(R(U)|XY = a!a)|$$

soit pour les mêmes raisons que dans la démonstration du théorème I-3 :

$$|R(U)|X = Y| = |(U\bar{R}(X)) \cap (U\bar{R}(Y))| \times E|(R(U)|XY = a!a)|$$

2) $R(U)|X > Y$, $XY \subseteq U$

$$|R(U)|X > Y| = \sum_{y \in U\bar{R}(Y)} \sum_{\substack{x \in U\bar{R}(X) \\ \text{et } x>y}} |(R(U)|XY = xy)|$$

soit :

$$|R(U)|X > Y| = |(R(U)|XY = xy)| \times \sum_{y \in U\bar{R}(Y)} n_y$$

où $n_y = |\{x \in U\bar{R}(X)/x > y\}|$

Le troisième cas est symétrique du second.

Théorème I-5

P est valide dans CE pour l'opérateur d'intersection. □

Démonstration :

Cet opérateur est un cas particulier de l'opérateur de join. Aussi nous n'allons pas étudier tous les cas possibles ici, nous renvoyons le lecteur à la démonstration du théorème I-8.

Néanmoins, nous allons donner une démonstration de ce résultat dans le cas où il y a indépendance intra-relationnelle dans les deux relations.

Soit $\bar{R}'(U)$ et $\bar{R}''(U)$ les ensembles résultat de l'union des référentiels de R'(U) et R"(U). Soit $\bar{R}(U) = \bar{R}'(U) \cap \bar{R}''(U)$.

Alors :

$$E|R(U)| = \sum_{u \in \bar{R}(U)} \text{prob}(u \in R'(U) \text{ et } u \in R''(U))$$

On sait que les référentiels d'une même relation sont disjoints, donc :

$\forall\, u \in R(U)$, $\exists!\, i$ et $\exists!\, j$ tels que $\bar{R}'_i(U) \cap \bar{R}''_j(U)$ où $\bar{R}'_i(U)$ et $\bar{R}''_j(U)$ sont deux référentiels de R'(U) et R"(U).

Ici trois cas se présentent :

1) Soit $u \in R'(U)$ est indépendant de $u \in R''(U)$

2) Soit $R'(U) \cap \bar{R}''_j(U) \cap \bar{R}'_i(U) \subseteq R''(U) \cap \bar{R}'_i(U) \cap \bar{R}''_j(U)$

3) Soit $R'(U) \cap \bar{R}''_j(U) \cap \bar{R}'_i(U) \supseteq R''(U) \cap \bar{R}'_i(U) \cap \bar{R}''_j(U)$

$\forall\, u \in (\bar{R}'_i \cap \bar{R}''_j)(U)$

Dans le premier cas :

$$\text{prob}(u \in R'(U) \text{ et } u \in R''(U)) = \text{prob}(u \in R'(U)) \times \text{prob}(u \in R''(U))$$

soit :

$$\text{prob}(u \in R'(U) \text{ et } u \in R''(U)) = \frac{E|R'(U)|}{|\bar{R}'(U)|} \times \frac{E|R''(U)|}{|\bar{R}''(U)|}$$

Dans le deuxième cas :

$$\text{prob}(u \in R'(U) \text{ et } u \in R''(U)) = \frac{E|R'(U)|}{|\bar{R}'(U)|}$$

Le troisième cas est symétrique du second.

Pour trouver le résultat final, il suffit de calculer chacune des espérances pour chaque référentiel correspondant à chacun de ces cas.

Soit pour le premier cas :

$$E|R(U) \cap (\bar{R}'_i(U) \cap \bar{R}''_j(U))| = |\bar{R}'_i(U) \cap \bar{R}''_j(U)| \times \frac{E|R'(U)| \times E|R''(U)|}{|\bar{R}'(U)| \times |\bar{R}''(U)|}$$

Dans le deuxième cas :

$$E|R(U) \cap (\bar{R}'_i(U) \cap \bar{R}''_j(U))| = |\bar{R}'_i(U) \cap \bar{R}''_j(U)| \times \frac{E|R'(U)|}{|\bar{R}'(U)|}$$

Le troisième cas est évidemment le symétrique du second.

Il reste à déterminer les espérances des tailles des projections de R(U). Ce calcul est un cas particulier du calcul effectué pour le join.

En effet, $R'(U) \cap R''(U) = R'[U = U]R'' = R(U)$.

Donc $\forall\ V \subsetneq U$, déterminer $E|R(V)|$, pour l'intersection est équivalent à le déterminer pour le join.

Théorème I-6

P est valide pour les opérateurs d'union et de différence dans CE. □

Démonstration :

Soit $S \in CE$ et $R(U) = R'(U) \cup R''(U)$ où $R'(U)$ et $R''(U)$ apppartiennent à S.

$$E|R(U)| = E|R'(U)| + E|R''(U)| - E|R'(U) \cap R''(U)|.$$

de même si $R(U) = R'(U) - R''(U)$, $E|R(U)| = E|R'(U)| - E|R'(U) \cap R''(U)|$

Pour diminuer le nombre de cas à étudier pour démontrer la validité du système pour l'équi-join, nous allons d'abors démontrer le théorème suivant.

Théorème I-7

Soient R(XY) et R'(XZ) deux relations et R"(XYZ) leur join sur X.

$\forall V \subset XYZ \;/\; X \subseteq V$ alors :

$R''(XYZ)\,[V] = R(XY \cap V) * R'(XZ \cap Y)$. □

Démonstration :

Nous montrons d'abord que :

$R''(XYZ)\,[V] \subseteq R(XY \cap V)\,[X = X]\,R'(XZ \cap V)$

Posons $V = Y_1Z_1X$, $Y = Y_1Y_2$ et $Z = Z_1Z_2$

1) Cela revient à

$R''(XYZ)\,[XY_1Z_1] \subseteq R(XY_1)\,[X = X]\,R'(XZ_1)$

soit $xy_1z_1 \in R''(XYZ)\,[XY_1Z_1]$

par définition de la projection, $\exists y_2z_2$ tels que

$xy_1y_2z_1z_2 \in R''(XYZ)$

par définition du join :

$xy_1y_2 \in R(XY)$ et $xz_1z_2 \in R(XZ)$

donc $xy_1 \in R(XY)\,[Y_1X]$ et $xz_1 \in R'(XZ)\,[Z_1X]$

donc par définition du join

$xy_1z_1 \in (R(XY)\,[XY_1])\,[X = X]\,(R'(XZ)\,[XZ_1])$

2) $R''(XYZ)\ [XY_1Z_1] \supseteq R(XY_1)\ [X = X]\ R'(XZ_1)$

si $xy_1z_1 \in R(XY_1)\ [X = X]\ R'(XZ_1)$

alors : $xy_1 \in R(XY_1)$ et $xz_1 \in R'(XZ_1)$

donc $\exists\, y_2z_2$ tels que

$xy_1y_2 \in R(XY)$ et $xz_1z_2 \in R'(XZ)$

donc par définition du join :

$xy_1y_2z_1z_2 \in R''(XYZ)$

Par définition de la projection :

$xy_1z_1 \in R''(XYZ)\ [XY_1Z_1]$.

Théorème I-8

P est valide dans CE pour l'opération d'équi-join. □

Démonstration :

Nous allons donner une démonstration directe de ce résultat.

Soit le join : $R''(XYZ) = R(XY)\ [X = X]\ R'(XZ)$

Pour montrer la validité du système de paramètres pour cette opération, il suffit d'après le théorème I-7 de calculer :

a) $E|R''(XYZ)|$

b) $E|R''(X'YZ)|\quad \forall\ X' \subset XYZ \quad X' \neq X$.

Nous notons $\bar{R}(Y)$ (resp $\bar{R}(Z), \bar{R}(X), \bar{R}(X')$) l'union de tous les référentiels de $R(Y)$ (resp $R'(Z), \bar{R}(X), \bar{R}(X')$).

Nous supposons d'abord que :

$\forall\ (x,x') \in \bar{R}(X)$ (resp $\bar{R}'(X)$) :

$x \in R(X)$ (resp $R'(X)$) est indépendant de $x' \in R(x)$ (resp $R'(X)$) (condition d'indépendance intra-relationnelle).

Nous calculons d'abord a).

$$P[|R''(XYZ)| = k] = \sum_{n=0}^{\infty} \text{prob}[|R(X) \cap R'(X)| = n] \times \text{prob}[|\bigsqcup_{i=1}^{n} R(x_i,XY) \times R(x_i,XZ)| = k$$

d'où :

$$E|R''(XYZ)| = \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} k \; prob[|R(X)\cap R'(X)| = n] \times prob[|\bigcup_{i=1}^{n} R(x_i,XY)\times R'(x_i,XZ)| = k]$$

$$E|R''(XYZ)| = \sum_{n=0}^{\infty} prob[|R(X)\cap R'(X)| = n] \times \sum_{k=0}^{\infty} k \; prob[|\bigcup_{i=1}^{n} R(x_i,XY)\times R'(x_i,XZ)| = k]$$

$$= \sum_{n=0}^{\infty} prob[|R(X)\cap R'(X)| = n] \times E\left(|\bigcup_{i=1}^{n} R(x_i,XY)\times R'(x_i,XZ)|\right)$$

Nous allons d'abord estimer le premier terme du produit.

- soit $prob[|R(X)\cap R'(X)| = n]$

  Ce terme dépend des référentiels dans lesquels se trouvent les valeurs de l'intersection. A ce stade, il faut répartir cette somme en sommes locales prises pour chaque couples de référentiels de R(X) et R'(X).

  Soit $\bar{R}_1(X)$ et $\bar{R}'_1(X)$ deux quelconques de ces référentiels, nous allons calculer :

  $E|R''(XYZ) \cap (\bar{R}_1(X) \times \bar{R}(Y)) \cap (\bar{R}'_1(X) \times \bar{R}'(Z))|$

  Pour trouver le résultat final , il suffit de faire la somme des espérances pour tous les couples $\bar{R}_1(X)$, $\bar{R}'_1(X)$ dont l'intersection n'est pas vide. En effet, par définition des référentiels :

$$[R''(XYZ)\cap(\bar{R}_1(X)\cap\bar{R}'_1(X))\times(\bar{R}(Y)\times\bar{R}'(Z))] \cap [R''(XYZ)\cap(\bar{R}_2(X)\cap\bar{R}'_2(X))\times(\bar{R}(Y)\times\bar{R}'(Z))] = \emptyset$$

  (En effet, $\bar{R}_1(X) \cap \bar{R}_2(X) = \emptyset$).

  1er cas : $R(X) \cap \bar{R}'_1(X) \cap \bar{R}_1(X) \subseteq R'(X) \cap \bar{R}_1(X) \cap \bar{R}'_1(X)$

  ici, $\forall x \in \bar{R}_1(X) \cap \bar{R}'_1(X)$:

  $$P[x \in R(X) \text{ et } x \in R'(X)] = \frac{E|R(X)|}{|\bar{R}(X)|}$$

  et d'après l'hypothèse (indépendance des n-uplets) :

$$\text{prob}[|R(X) \cap R'(X) \cap \bar{R}_1(X) \cap \bar{R}'_1(X)| = k]$$

$$= C^{k}_{|\bar{R}_1(X)\cap\bar{R}'_1(X)|} \left(\frac{E|R(X)|}{|\bar{R}(X)|}\right)^{k} \left(1 - \frac{E|R(X)|}{|\bar{R}(X)|}\right)^{|\bar{R}_1(X)\cap\bar{R}'_1(X)|}$$

2ème cas : $\forall\ x \in \bar{R}_1(X) \cap \bar{R}'_1(X)$, $x \in \bar{R}(X)$ est indépendant de $x \in \bar{R}'(X)$.

d'où :

$$\text{prob}[x \in R(X) \text{ et } x \in R'(X)] = \frac{E|R(X)| \times E|R'(X)|}{|\bar{R}(X)| \times |\bar{R}'(X)|}$$

et d'après l'hypothèse d'indépendance intra-relationnelle :

$$\text{prob}[|R(X) \cap R'(X) \cap \bar{R}_1(X) \cap \bar{R}'_1(X)| = k]$$

$$= C^{k}_{|\bar{R}_1(X)\cap\bar{R}'_1(X)|} \left(\frac{E|R(X)|\times E|R'(X)|}{|\bar{R}(X)| \times |\bar{R}'(X)|}\right)^{k} \left(1 - \frac{E|R(X)|\times E|R'(X)|}{|\bar{R}(X)| \times |\bar{R}'(X)|}\right)^{|R_1(X)\cap\bar{R}'_1(X)|-k}$$

Dans ce qui suit, pour simplifier les notations, nous noterons :

$$\text{prob}[|R(X) \cap R'(X) \cap \bar{R}_1(X) \cap \bar{R}'_1(X)| = k] = P_k$$

- Il faut maintenant calculer $E\left(\left|\bigsqcup_{i=1}^{n} R(x_i,XY) \times R'(x_i,XZ)\right|\right)$

Par définition de $R(x_i,XY)$ et de $R'(x_i,XZ)$ nous avons :

$\forall\ i,j\ ;\ i \neq j$:

$(R(x_i,XY) \times R'(x_i,XZ)) \cap (R(x_j,XY) \times R'(x_j,XZ)) = \emptyset$

donc : $\left|\bigsqcup_{i=1}^{n} R(x_i,XY) \times R'(x_i,XZ)\right| = \sum_{i=1}^{n} |R(x_i,XY) \times R'(x_i,XZ)|$

soit par passage à l'espérance :

$$E\left|\bigsqcup_{i=1}^{n} R(x_i,XY) \times R'(x_i,XZ)\right| = \sum_{i=1}^{n} E|R(x_i,XY) \times R'(x_i,XZ)|$$

Nous avons :

$$E|R(x_i,XY)| = \frac{E|R(XY)|}{E|R(X)|} \quad \text{et} \quad E|R'(x_i,XZ)| = \frac{E|R'(XZ)|}{E|R'(X)|}$$

D'après les propriétés de la classe CE, nous avons :

$$E|R(x_i,XY) \times R'(x_i,XZ)| = \frac{E|R(XY)| \times E|R'(XZ)|}{E|R(X)| \times E|R'(X)|}$$

d'où finalement :

$$E|R"(XYZ) \cap [(\bar{R}_1(X) \cap \bar{R}'_1(X)) \times \bar{R}(Y) \times \bar{R}'(Z)]| = \sum_{k=0}^{\infty} P_k \times k \times \frac{E|R(XY)| \times E|R'(XZ)|}{E|R(X)| \times E|R'(X)|}$$

* soit dans le cas où $x \in R(X)$ est indépendant de $x \in R'(X)$

$$E|R"(XYZ) \cap [(\bar{R}_1(X) \cap \bar{R}'_1(X)) \times \bar{R}(Y) \times \bar{R}'(Z)]| = E|R(XY)| \times E|R'(XZ)| \times \frac{|\bar{R}_1(X) \cap \bar{R}'_1(X)|}{|\bar{R}(X)| \times |\bar{R}'(X)|}$$

* soit dans le cas où $R(X) \cap \bar{R}^{-1}(X) \subseteq R'(X) \cap \bar{R}(X)$

$$E|R"(XYZ) \cap [(\bar{R}_1(X) \cap \bar{R}'_1(X)) \times \bar{R}(Y) \times \bar{R}'(Z)]| = \frac{E|R(XY)| \times E|R'(XZ)| \times |\bar{R}_1(X) \cap \bar{R}'_1(X)|}{|\bar{R}(X)| \times E|R'(X)|}$$

Toujours dans l'hypothèse générale que nous avons faite (i.e. indépendance intra-relationnelle dans chaque relation), nous allons évaluer $E|R(V) \cap [(\bar{R}_1(X) \cap \bar{R}'_1(X)) \times \bar{R}(Y) \times \bar{R}'(Z)]|$ dans le cas où $X \neq V$

Soit $X' = X \cap V$ et $V = X'YZ$.

Le calcul de P_k reste le même, par contre il faut calculer différement :

$$E|\bigcup_{i=1}^{n} R(x_i,X'Y) \times R'(x_i,X'Z)|$$

Ici les relations $R(x_i,X'Y)$ et $R(x_j,X'Y)$ ne sont pas forcément disjointes.

Nous avons :

$$\left|\bigcup_{i=1}^{n} (R(x_i,X'Y) \times R'(x_i,X'Z)\right| = \left|\bigcup_{i=1}^{n} (R''(x_i,X'YZ))\right| =$$

$$|R''(x_1,X'YZ)| + |R''(x_2,X'YZ)| - |R''(x_1,X'YZ) \cap R''(x_2,X'YZ)|$$

$$+ |R''(x_3,X'YZ)| - |R''(x_3,X'YZ) \cap (R(x_1,XYZ) \cup R(x_2,XYZ))| + \ldots$$

Comme précédemment, nous avons :

$$E|R''(x_i,X'YZ)| = \frac{E|R(XY)| \times E|R'(XZ)|}{E|R(X)| \times E|R'(X)|}$$

Pour simplifier les notations, nous noterons cette espérance $E(X'YZ)$.

Pour terminer il faut déterminer l'espérance de la taille des intersections des relations présentes dans la formule précédente.

1er cas : $\exists\, V \subseteq X'Y$ ou $\subseteq X'Z$ / $V \rightarrow X$

Dans cette hypothèse

$\forall\, i,j,\ R''(x_i,X'YZ) \cap R''(x_j,X'YZ) = \emptyset$

Nous sommes ramenés au calcul de $E|R''(XYZ)|$

2ème cas : $\forall\, V \subseteq X'Y$ et $\subseteq X'Z$ / $V \not\rightarrow X$

Dans ce qui suit, par abus de langage, pour simplifier, nous notons $E|R''(XYZ)\ [X'YZ]|$ par $E|R''[X'YZ]|$.

D'après les hypothèses, nous avons :

$$E(|R''(x,X'YZ) \cap R'(x',X'YZ)|) = \frac{E|R''(x,X'YZ)| \times E|R''(x',X'YZ)| \times |\bar{R}_1''(X'YZ)|}{|\bar{R}''(X'YZ)| \times |\bar{R}''(X'YZ)|}$$

où $\bar{R}''(X'YZ) = (\bar{R}(X') \cap \bar{R}'(X')) \times \bar{R}(Y) \times \bar{R}'(Z)$

et $\bar{R}_1''(X'YZ) = (\bar{R}_1(X') \cap \bar{R}_1'(X')) \times \bar{R}(Y) \times \bar{R}'(Z)$

d'où :

$$E\left|\bigcup_{i=1}^{n} R''(x_i, X'YZ)\right| = E(X'YZ) + E(X'YZ) - \frac{E(X'YZ)^2 \times |\bar{R}_1''(X'YZ)|}{|\bar{R}''(X'YZ)|^2}$$

$$+ E(X'YZ) - \left[2E(X'YZ) - \frac{E(X'YZ)^2 \times |\bar{R}_1''(X'YZ)|}{|\bar{R}''(X'YZ)|^2}\right]\left(\frac{E(X'YZ)}{|\bar{R}''(X'YZ)|^2}\right) + \ldots$$

Nous nous trouvons devant une suite géométrique de raison $\left(1 - \frac{E(X'YZ) \times |\bar{R}_1''(X'YZ)|}{|\bar{R}''(X'YZ)|^2}\right)$ et de premier terme $E(X'YZ)$.

D'où le calcul du kième terme qui est :

$$E\left|\bigcup_{i=1}^{k} R''(x_i, X'YZ)\right| = E(X'YZ) \times \frac{1 - \left(1 - \frac{E(X'YZ) \times |\bar{R}_1''(X'YZ)|}{|\bar{R}''(XYZ)|^2}\right)^k}{1 - \left(1 - \frac{E(X'YZ) \times |\bar{R}_1''(X'YZ)|}{|\bar{R}''(XYZ)|^2}\right)}$$

Nous pouvons calculer le résultat final :

$$E|R''[X'YZ] \cap \bar{R}_1''(X'YZ)| = \sum_{k=0}^{\infty} E(X'YZ) \frac{1 - \left(1 - \frac{E(X'YZ) \times |\bar{R}_1''(X'YZ)|}{|\bar{R}''(X'YZ)|^2}\right)^k}{1 - \left(1 - \frac{E(X'YZ) \times |\bar{R}_1''(X'YZ)|}{|\bar{R}''(X'YZ)|^2}\right)} \times P_k$$

Pour simplifier le calcul de cette formule, nous posons :

$$\alpha = 1 - \frac{E(X'YZ) \times |\bar{R}_1''(X'YZ)|}{|\bar{R}''(X'YZ)|^2}$$

d'où maintenant :

$$E|R''[X'YZ] \cap \bar{R}_1''(X'YZ)| = \sum_{k=0}^{\infty} E(X'YZ) \times \frac{1 - \alpha^k}{1 - \alpha} \times P_k$$

$$= \frac{E(X'YZ)}{\alpha - 1} \times \sum_{k=0}^{\infty} (\alpha^k - 1) P_k = -\frac{E(X'YZ)}{\alpha - 1} + \left(\sum_{k=0}^{\infty} \alpha^k P_k\right) \times \frac{E(X'YZ)}{\alpha - 1}$$

P_k est de la forme $C_\gamma^k (p)^k (1 - p)^{\gamma-k}$

d'où :

$$E|R"[X'YZ] \cap \bar{R}_1"(X'YZ)| = - \frac{E(X'YZ)}{\alpha - 1} + \left(\sum_{k=0}^{\gamma} C_\gamma^k (p)^k (1 - p)^{\gamma-k} \right) \times \frac{E(X'YZ)}{\alpha - 1}$$

$$E|R"[X'YZ] \cap \bar{R}_1"(X'YZ)| = - \frac{E(X'YZ)}{\alpha - 1} + \frac{E(X'YZ)}{\alpha - 1} (1 - p + p\alpha)^\gamma$$

$$= \frac{E(X'YZ)}{-\alpha + 1} (1 - (1 - p(\alpha - 1)^\gamma) \qquad (*)$$

Pour obtenir le résultat définitif il faut remplacer α par $1 - \frac{E(X'YZ) \times |\bar{R}_1"(X'YZ)|}{|\bar{R}"(X'YZ)|^2}$ et distinguer deux cas pour évaluer les p et γ dans la probabilité P_k, suivant la condition 2 (i.e. indépendance intra-relationnelle ou non).

1er cas : indépendance intra-relationnelle entre R(X) et R'(X) pour $\bar{R}_1"(X)$:

$$E|R"[X'YZ] \cap \bar{R}_1"(X'YZ)| =$$

$$\frac{|\bar{R}"(X'YZ)|^2}{|\bar{R}_1"(X'YZ)|} \left(1 - \left(1 - \frac{E(X'YZ) \times |\bar{R}_1"(X'YZ)|}{|\bar{R}"(X'YZ)|^2} \times \frac{E|R(X)| \times E|R'(X)|}{|\bar{R}(X)| \times |\bar{R}'(X)|} \right)^{|\bar{R}_1(X) \cap \bar{R}_1'(X)} \right.$$

d'où en remplaçant E(X'YZ) par sa valeur :

$$E|R"[X'YZ] \cap \bar{R}_1"(X'YZ)| =$$

$$\frac{|\bar{R}"(X'YZ)|^2}{|\bar{R}_1"(X'YZ)|} \left(1 - \left(1 - \frac{E|R(XY)| \times E|R'(XZ)| \times |\bar{R}_1"(X'YZ)|}{|\bar{R}"(X'YZ)|^2 \times |\bar{R}(X)| \times |\bar{R}'(X)|} \right)^{|\bar{R}_1(X) \cap \bar{R}_1'(X)|} \right)$$

2ème cas : $R(X) \cap \bar{R}_1"(X) \subseteq R'(X) \cap \bar{R}_1"(X)$, il suffit de changer la valeur de P_k dans la formule, soit :

$$E|R"[X'YZ] \cap \bar{R}_1"(X'YZ)| =$$

$$\frac{|\bar{R}"(X'YZ)|^2}{|\bar{R}"(X'YZ)|} \left(1 - \left(1 - \frac{E(X'YZ) \times |\bar{R}_1"(X'YZ)|}{|\bar{R}"(X'YZ)|^2} \times \frac{E|R(X)|}{|\bar{R}(X)|} \right)^{|\bar{R}_1(X) \cap \bar{R}_1'(X)|} \right)$$

d'où en remplaçant E(X'YZ) par sa valeur :

$$E|R''[X'YZ] \cap \bar{R}_1''(X'YZ)|$$

$$= \frac{|\bar{R}''(X'YZ)|^2}{|\bar{R}_1''(X'YZ)|} \left(1 - \left(1 - \frac{E|R(XY)| \times E|R'(XZ)| \times |\bar{R}_1''(X'YZ)|}{|\bar{R}''(X'YZ)|^2 \times |\bar{R}(X)| \times E|R'(X)|}\right)^{|\bar{R}_1(X) \cap \bar{R}_1'(X)|}\right)$$

Nous avonc ici calculé l'espérance de la taille de R"(V) dans le cas où V contient ou ne contient pas les attributs sur lesquels s'effectue l'équi-join, ceci en étudiant les deux cas possibles de la condition 2. Cependant, nous nous sommes placés dans un cas de la condition 3 : en effet, nous avons supposé qu'il y avait indépendance intra-relationnelle dans les deux relations sur l'ensemble d'attributs sur lequel s'opère l'équi-join. Il reste donc à effectuer les calculs pour les cas où l'on n'a plus indépendance des n-uplets sur cet ensemble dans l'une des deux relations ou dans les deux.

Dans ce cas, d'après la propriété de la classe CE, il existe une partition (U,V) de X telle que $(U \rightarrow V)$ est une dépendance fonctionnelle dans R(X) ou R'(X). Comme l'intersection de R(X) et R'(X) est un sous ensemble de chacune des deux relations, il est clair que la dépendance fonctionnelle $(U \rightarrow V)$ sera vérifiée pour l'intersection. La seule modification intervenant par rapport aux calculs précédents est dans l'évaluation de la probabilité $prob|R(X) \cap R'(X) = k|$. Nous allons maintenant évaluer cette probabilité (nous gardons les mêmes notations que précédemment et les mêmes hypothèses sur les référentiels).

Soient R(UV) et R'(UV) les deux projections sur lesquelles se fait le test de l'égalité dans l'équi-join (i.e. UV est une partition de X). Soient $\bar{R}_1''(U)$ (resp. $\bar{R}_1''(V)$) l'intersection de $\bar{R}_1(U)$ et $\bar{R}_1'(U)$ (resp. $\bar{R}_1(V)$ et $\bar{R}_1'(V)$).

Nous nous plaçons ici dans le premier cas de la condition 2 (les autres cas se traitent comme ceux de la première partie).

Soit uv $\in$ R(UV) est indépendant de uv $\in$ R'(UV).

$|R(UV) \cap R'(UV)| = n$ est équivalent à dire que : il existe n couples (U,V) dans $\bar{R}''_1(U) \times \bar{R}''_1(V)$ du type $(u_1,v),\ldots,(u_n,v)$ tels que :

1) $\forall\ 1 \leq i,j \leq n,\ i \neq j,\ u_i \neq u_j$

2) v est quelconque dans chaque couple.

et $\forall\ u_k,\ k \notin 1,\ldots,n$ et $\forall\ v \in \bar{R}''_1(V)$: $(u_k,v) \notin R(UV) \cap R'(UV)$.

Il reste à traduire cela en termes de probabilités.

Il y a $\binom{n}{|\bar{R}''_1(U)|}$ façons de choisir les éléments de $\bar{R}''_1(U)$ pour constituer n couples.

Il y a $|\bar{R}''_1(V)|$ choix possibles pour v.

donc :

$$\text{prob}(|R(UV)\cap R'(UV)| = n) = \binom{n}{|\bar{R}''_1(U)|} \text{prob}((u_1,v) \text{ et} \ldots \text{et } (u_n,v) \in R(UV) \text{ et } R'(UV) \forall\ v \in \bar{R}''_1(V)) \times \text{prob}(\forall\ k \notin 1,\ldots,n,\ \forall\ v,\ (u_k,v) \notin R(UV) \cap R'(UV))$$

Il reste à évaluer les deux probabilités ci-dessus.

$$\text{prob}((u_1,v) \text{ et } \ldots \text{ et } (u_n,v) \in R(UV) \text{ et } R'(UV),\ \forall\ v \in \bar{R}''_1(V))$$
$$= \text{prob}((u_1,v) \in R(UV)\cap R'(UV),\ \forall\ v \in \bar{R}''_1(V)) \times \ldots \times \text{prob}((u_n,v) \in R(UV)\cap R'(UV),\ \forall\ v \in \bar{R}''_1(V))$$

En effet, si la suite de couples vérifie l'affirmation 1, les événements sont indépendants.

$$\text{prob}((u,v) \in R(UV)\cap R'(UV),\ \forall\ v \in \bar{R}''_1(V)) = \left(\frac{|\bar{R}''_1(V)| \times E|R(UV)| \times E|R'(UV)|}{|\bar{R}(UV)| \times |\bar{R}'(UV)|} \right)$$

Il reste à évaluer :

$$\text{prob}((u_k,v) \notin R(UV) \cap R'(UV),\ \forall\ v \in \bar{R}''_1(V),\ \forall\ k > n)$$

Ici encore, nous avons :

$$\text{prob}((u_k,v) \notin R(UV) \cap R'(UV),\ \forall\ v \in \bar{R}''_1(V),\ \forall\ k > n)$$

$$= \text{prob}((u_{n+1},v) \text{ et } (u_{n+2},v) \text{ et } \ldots \text{ et } (u_{|\bar{R}''_1(U)|},v) \notin R(UV) \cap R'(UV),\ \forall\ v \in \bar{R}''_1(V))$$

$$= \text{prob}((u_{n+1},v) \notin R(UV) \cap R'(UV),\ \forall\ v \in \bar{R}''_1(V)) \times \ldots \times \text{prob}((u_{|\bar{R}''_1(U)|},v) \notin R(UV) \cap R'(UV),\ \forall\ v \in \bar{R}''_1(V)).$$

En effet, là aussi, les probabilités sont indépendantes.

$$\text{prob}((u_i,v) \notin R(UV) \cap R'(UV),\ \forall\ v \in \bar{R}''_1(V)) = \left(1 - \frac{|\bar{R}''_1(V)| \times E|R(UV)| \times E|R'(UV)|}{|\bar{R}(UV)| \times |\bar{R}'(UV)|}\right)$$

d'où

$$\text{prob}(|R(UV) \cap R'(UV)| = |R(X) \cap R'(X)| = n)$$

$$= C^{n}_{|\bar{R}''_1(U)|} \left(\frac{|\bar{R}''_1(V)| \times E|R(UV)| \times E|R'(UV)|}{|\bar{R}(UV)| \times |\bar{R}'(UV)|}\right)^{n} \times \left(1 - \frac{|\bar{R}''_1(V)| \times E|R(UV)| \times E|R'(UV)|}{|\bar{R}(UV)| \times |\bar{R}'(UV)|}\right)^{|\bar{R}''_1(U)|-n}$$

Remarque : dans l'autre cas de la condition 2, $R(UV) \cap \bar{R}'_1(UV) \subseteq R'(UV) \cap \bar{R}_1(UV)$

La seule modification est :

$$\text{prob}((u,v) \in R(UV) \cap R'(UV),\ \forall\ v \in \bar{R}''_1(V)) = \left(\frac{|\bar{R}''_1(V)| \times E|R(UV)|}{|\bar{R}(UV)|}\right)$$

Le reste des calculs est identique. Nous citons ce cas comme référence, toutes les autres possibilités se traitant de manière similaire (par exemple : $R(U) \cap \bar{R}'_1(U) \subseteq R'(U) \cap \bar{R}_1(U)$, etc...)

Nous allons maintenant appliquer ce résultat au calcul de l'espérance de la taille d'un équi-join.

Nous rappelons que $X = UV$ et que $U \rightarrow V$ dans $R(XY)$ ou dans $R'(XY)$.

Nous allons d'abord estimer $E|R''(XYZ) \cap \bar{R}''_1(XYZ)|$

$$E|\bar{R}''(XYZ)\cap\bar{R}_1''(XYZ)| = \sum_{n=0}^{\infty} \text{prob}(|R(X)\cap R'(X)| = n) \times \frac{E|R(XY)|\times E|R'(XZ)|}{E|R(X)|\times E|R'(X)|} \times n$$

$$= \sum_{n=0}^{\infty} C^{n}_{|\bar{R}_1''(U)|} \left(\frac{|\bar{R}_1''(V)|\times E|R(UV)|\times E|R'(UV)|}{|\bar{R}(UV)|\times|\bar{R}'(UV)|}\right)^{n} \left(1 - \frac{|\bar{R}_1''(V)|\times E|R(UV)|\times E|R'(UV)|}{|\bar{R}(UV)|\times|\bar{R}'(UV)|}\right)^{|\bar{R}_1''(U)|-}$$

$$\times \frac{E|R(XY)|\times E|R'(XZ)|}{E|R(X)|\times E|R'(X)|} \times n$$

$$= \frac{|\bar{R}_1''(U)|\times|\bar{R}_1''(V)|\times E|R(UV)|\times E|R'(UV)|}{|\bar{R}(UV)|\times|\bar{R}'(UV)|} \times \frac{E|R(XY)|\times E|R'(XZ)|}{E|R(X)|\times E|R'(X)|}$$

$$\text{soit } \frac{|\bar{R}_1''(U)| \times |\bar{R}_1''(V)| \times E|R(XY)| \times E|R'(XZ)|}{|\bar{R}(UV)| \times |\bar{R}'(UV)|}$$

$$\text{soit sachant que } UV = X : \frac{|\bar{R}_1''(X)| \times E|R(XY)| \times E|R'(XZ)|}{|\bar{R}(X)| \times |\bar{R}'(X)|}$$

Il reste à effectuer le calcul pour déterminer $R''[X'YZ]$ où $X' \subset X$.

Pour cela, il suffit de reprendre la formule (*) obtenue précédemment précédemment (page **106**) et remplacer P_k par la nouvelle valeur. Nous nous plaçons dans la condition 2 (indépendance intra-relationnelle entre R(X) et R'(X)).

$$E|R''[X'YZ]\cap\bar{R}_1''(X'YZ)| = \frac{|R''[X'YZ]|^2}{|\bar{R}_1''(X'YZ)|} \left[1 - \left[1 - \frac{|\bar{R}_1''(V)|\times E|R(UV)|\times E|R'(UV)|}{|\bar{R}(UV)|\times|\bar{R}'(UV)|}\right.\right.$$

$$\left.\left. + \frac{|\bar{R}_1''(V)|\times E|R(UV)|\times E|R'(UV)|}{|\bar{R}(UV)|\times|\bar{R}'(UV)|} \left(1 - \frac{E(X'YZ)\times|\bar{R}_1''(X'YZ)|}{|\bar{R}''(X'YZ)|^2}\right)\right]^{|\bar{R}_1''(U)|}\right]$$

d'où en remplaçant E(X'YZ) par sa valeur :

$$= \frac{|\bar{R}''[X'YZ]|^2}{|\bar{R}_1''(X'YZ)|} \left[1 - \left[1 - \frac{|\bar{R}_1''(V)| \times E|R(XY)| \times E|R'(XZ)| \times |\bar{R}_1''(X'YZ)|}{|\bar{R}(X)| \times |\bar{R}'(X)| \times |\bar{R}''(X'YZ)|^2} \right]^{|\bar{R}_1''(U)|} \right]$$

Le cas où $R(UV) \cap \bar{R}'(UV) \subseteq R'(UV) \cap \bar{R}(UV)$ se traite de façon similaire.

Théorème I-9

P est valide dans CE pour l'opération de join généralisé. □

Démonstration :

D'après le Lemme VI-3 : $X \subseteq U$, $Y \subseteq V$

$R(U)\ [X \ominus X]\ R'(V) = (R(U) \otimes R'(V))\ |\ X \ominus Y$

Tout join généralisé est donc un produit cartésien suivi d'une sélection. Il suffit donc de montrer que P est valide dans CE pour le produit cartésien.

1) $E|R(U) \otimes R'(V)| = E|R(U)| \otimes E|R'(V)|$

2) $\forall\ W \subseteq UV$

$E|(R(U) \otimes R'(V))\ [V]| = E|(R(U)\ [U \cap W])| \otimes E|(R'(U)\ [V \cap W])$

Nous avons montré de manière constructive que le système P est valide dans CE pour toutes les opérations de l'algèbre relationnelle. Nous avons donc un outil nous permettant d'évaluer l'espérance de la taille des relations construites à partir d'un schéma de la classe CE par n'importe quelle requête exprimée dans l'algèbre $\mathcal{R}$.

Nous montrons maintenant que le système P_A défini en II.2 dans le chapitre II est aussi valide dans la classe CE.

Théorème I-10

P_A est valide dans CE pour l'algèbre $\mathcal{A}$. □

Démonstration :

Pour démontrer ce résultat, il suffit de montrer que P_A permet de calculer pour toute relation les espérances des tailles des projections de cette relation. C'est un résultat montré par Demolombe [Demo 81], qui constitue en fait un cas particulier de la démonstration de la validité de P pour l'intersection.

Soit R(U) une relation de R et $X \subset U$.

Si $X = A_1,\ldots,A_n$ où les A_i sont des attributs simples.

$\forall\, i,\ E|R(A_i)| \in P_A(S)$ ainsi que $E|R(U)|$

Si nous supposons que les sous-n-uplets r(X) de R résultent de $E|R(U)|$ tirages aléatoires non exhaustifs dans $\prod_{i=1}^{n} E|R(A_i)|$ n-uplets possibles.

D'après le théorème de Jordan (voir le calcul de (R(XY) * R'(XZ)[X'YZ]), nous avons :

$$E|R(X)| = E|R(U)|\left(1 - \left(1 - \frac{1}{E|R(U)|}\right)^{\prod_{i=1}^{n} E|R(A_i)|}\right)$$

Si $\exists\ X', Y \subseteq X$ tels que $X' \rightarrow Y$ et $Y \cap X' = \emptyset$ alors :

$$E|R(X)| = E|R(U)|\left(1 - \left(1 - \frac{1}{E|R(U)|}\right)^{\bigcap_{A \in X-Y} E|R(A)|}\right)$$

La démonstration de la validité de P_A découle directement de celle de P. En effet, nous avons vu que P_A permet de déterminer les espérances des tailles des projections, nous sommes donc ramenés à un cas particulier du système P. □

Nous avons donné dans le chapitre II, la définition de deux autres systèmes de paramètres qui ont pour but de donner la distributions de taille des relations. Dans ce qui suit, nous allons montrer que l'un d'eux (le système P") est valide dans CE pour l'algèbre relationnelle.

II - VALIDITÉ DU SYSTÈME P" DANS LA CLASSE CE POUR LES OPÉRATIONS DE L'ALGÈBRE RELATIONNELLE

Nous allons d'abord rappeler la définition du système P".

Définition II-1 :

Nous notons P" le système de paramètres suivant :

Soit (S,p) un schéma probabilisé.

P" associe à S l'ensemble des distributions de toutes les projections des relations de S et une valeur de REF. □

Théorème II-1 :

$\forall$ (S,p) $\in$ CE

$\forall$ $R_t \in S$, $\forall$ $R_t(U) \in R_t$, $\forall$ $X,Y \subseteq U$

f_{X-Y} se construit à partir de $P"(R_t)$. □

Démonstration :

Nous rappelons que $f_{X-Y}(n)$ désigne la probabilité pour un élément x des référentiels de R(X) d'être associé à n éléments distincts des référentiels de R(Y).

Soit $\bar{R}(X)$ et $\bar{R}(Y)$ l'union des référentiels de R(X) et de R(Y).

Nous supposons que $Y \cap X = \emptyset$ (en effet si $Y = Y'X'$ et $X = X'X"$ alors $f_{X-Y}(n) = f_{X'X"-Y'}(n)$).

Si nous considérons x un élément de $\bar{R}(X)$ et y un élément de $\bar{R}(Y)$, nous avons d'après la condition 1 :

$$\text{prob}(xy \in R(XY)) = \frac{E|R(XY)|}{|\bar{R}(XY)|}$$

$$\text{or} \quad E|R(XY)| = \sum_{i=1}^{|\bar{R}(XY)|} i \; f_{\emptyset-XY}(i)$$

1) Nous nous plaçons d'abord dans le premier cas de la propriété 3 (indépendance intra-relationnelle) : $\forall$ xy $\in$ ($U\bar{R}(XY)$) :

xy $\in$ R(XY) est indépendant de xy $\in$ R'(XY).

Dans ce cas, chercher à déterminer $f_{X-Y}(n)$ $\forall$ n, revient à chercher à déterminer la probabilité d'avoir n couples $(xy)'$ dans R(XY), chacun ayant la même probabilité de s'y trouver. Ces couples sont choisis parmi {x} X $\bar{R}(Y)$

Nous nous trouvons devant une distribution binômiale :

$$f_{X-Y}(n) = C^{n}_{|U\bar{R}(Y)|} \left(\frac{E|R(XY)|}{|U\bar{R}(XY)|}\right)^{n} \times \left(1 - \frac{E|R(XY)|}{|U\bar{R}(XY)|}\right)^{|U\bar{R}(Y)| - n}$$

2) Nous nous plaçons maintenant dans le deuxième cas de la propriété 3 de la classe CE.

$\exists$ (U,V) une partition de (X,Y) telle que U $\rightarrow$ V dans R(XY). Nous avons différentes possibilités :

a) U = X et Y = V ;

dans ce cas, puisque X $\rightarrow$ Y dans R(XY) :

$f_{X-Y}(1) = 1$ et $f_{X-Y}(n) = 0$ $\quad \forall$ n $\neq$ 1

b) Y = Y'Y" où Y' = Y $\cap$ V ;

alors XY" $\rightarrow$ Y' puisque U $\subseteq$ XY"

donc $\forall$ n > $|U\bar{R}(Y")|$: $f_{X-Y}(n) = 0$

En effet, dans ce cas, le nombre de valeurs de Y pour X sera le nombre de valeurs de Y" pour X.

Soit $f_{X-Y}(n) = f_{X-Y"}(n)$

Nous savons que X $\not\rightarrow$ Y" dans R(XY) car alors Y" serait dans V ce qui contredirait l'hypothèse. Déterminer $f_{X-Y"}(n)$ revient donc à effectuer le calcul du 1) en remplaçant Y par Y".

Théorème II-2 :

P" est valide dans CE pour l'opérateur de projection. □

Démonstration :

La démonstration se déduit directement de la définition de P".

Théorème II-3 :

P" est valide dans CE pour l'opération de séléction élémentaire sur l'égalité (R(U) | X = x). □

Démonstration :

R'(U) = R(U) | X = x.

$\forall$ X' $\subseteq$ U, il faut déterminer $f'_{\emptyset\text{-}X'}(n)$ ($f'_{\emptyset\text{-}X'}(n)$ désigne la distribution des projections de R'(U)).

1) X' $\subseteq$ X

$$f'_{\emptyset\text{-}X'}(1) = \text{prob}(x \in R(U)) = \frac{E|R(X)|}{|\bar{R}(X)|} = \frac{\sum_{i=1}^{|R(X)|} i\, f_{\emptyset\text{-}X}(i)}{|U\bar{R}(X)|}$$

où $U\bar{R}(X)$ désigne l'union des référentiels de R(X).

2) X' $\subsetneq$ X

$$f'_{\emptyset\text{-}X'}(n) = f_{X\text{-}X'}(n)$$

En effet la distribution de taille de $R_1(U)$ [X'] est celle de R(x,X').

Théorème II-4 :

P" est valide dans CE pour l'opération d'intersection et l'opération de différence. □

Démonstration :

Soit l'opération suivante :

R(U) $\cap$ R'(U) = R"(U).

$\forall$ X $\subseteq$ U il faut déterminer $f''_{\emptyset\text{-}X}(n)$.

1) $X = U$: $\bar{R}(U)$ et $\bar{R}'(U)$ sont deux référentiels non disjoints de $R(U)$ et $R'(U)$.

$$(*)\quad f'(n) = \sum_{i=n}^{\infty} \sum_{j=n}^{\infty} f(i) \times f'(j) \times \text{prob}(|R(U) \cap R'(U)| = n \;/\; |R(U)| = i \text{ et } |R'(U)| = j)$$
$$\emptyset\text{-}U \qquad \emptyset\text{-}U \quad \emptyset\text{-}U$$

Il faut déterminer $\text{prob}(|R(U) \cap R'(U)| = n \;/\; |R(U)| = i \text{ et } |R'(U)| = j)$

A ce stade nous allons introduire les notations suivantes :

$$f(n)_{\emptyset\text{-}U|\bar{R}(U)} = \text{prob}(|\bar{R}(U) \cap R(U)| = n)$$

où $\bar{R}(U)$ est un référentiel de $R(U)$.

Nous allons donc déterminer $f^2(n)_{\emptyset\text{-}U|\bar{R}''(U)}$ où $\bar{R}''(U) = \bar{R}(U) \cap \bar{R}'(U)$.

Nous rappelons que $f(n)_{\emptyset\text{-}U}$ (resp. $f'(n)_{\emptyset\text{-}U}$) désigne la distribution de la taille de $R(U)$ (resp. $R'(U)$).

1er cas : indépendance intra-relationnelle dans $R(U)$ et $R'(U)$ et indépendance inter-relationnnelle de $R(U)$ et $R'(U)$ pour les éléments de $\bar{R}''(U)$.

$\text{prob}(|R(U) \cap R'(U) \cap \bar{R}''(U)| = n \;/\; |R(U)| = i \text{ et } |R'(U)| = j)$ suit une distribution binômiale.

$$\text{prob}(|R(U) \cap R'(U) \cap \bar{R}''(U)| = n \;/\; |R(U)| = i \text{ et } |R'(U)| = j)$$

$$= C^{n}_{|\bar{R}''(U)|} \left(\frac{ij}{|U\bar{R}(U)|\ |U\bar{R}'(U)|} \right)^{n} \left(1 - \frac{ij}{U\bar{R}(U)\quad U\bar{R}'(U)} \right)^{|\bar{R}''(U)| - n}$$

2ème cas : indépendance intra-relationnelle dans $R(U)$ et $R'(U)$ et $R(U) \cap \bar{R}''(U) \subseteq R'(U) \cap \bar{R}''(U)$.

$$\text{prob}(|R(U) \cap R'(U) \cap \bar{R}''(U)| = n \;/\; |R(U)| = i \text{ et } |R'(U)| = j) = \begin{cases} 0 \text{ si } n \neq i \\ 1 \text{ sinon} \end{cases}$$

3ème cas : indépendance inter-relationnelle de R(U) et R'(U).

Il existe (X,Y) une partition de U telle que $X \rightarrow Y$ dans R(U) ou dans R'(U).

Dans ce cas $X \rightarrow Y$ dans $R(U) \cap R'(U)$.

$$\text{prob}(|R(U) \cap R'(U) \cap \bar{R}''(U)| = n \;/\; |R(U)| = i \text{ et } |R'(U)| = j)$$

$$= C^{n}_{|\bar{R}''(X)|} \left(\frac{|\bar{R}''(Y)| \times i \times j}{|U\bar{R}(U)| \times |U\bar{R}'(U)|} \right)^{n} \left(1 - \frac{|\bar{R}''(Y)| \times i \times j}{|U\bar{R}(U)| \times |U\bar{R}'(U)|} \right)^{|\bar{R}''(X)| - n}$$

En remplaçant cette probabilité par sa valeur dans (*), suivant le cas, nous obtenons $f(n)_{\emptyset\text{-}U|\bar{R}''(U)}$

Le 4ème cas se traite comme le second.

Pour conclure, il faut déterminer cette distribution pour tous les couples de référentiels de R(U) et de R'(U) non disjoints.

Nous montrons que la connaissance de $f(n)_{\emptyset\text{-}U|\bar{R}(U)}$ $\forall\, \bar{R}(U)$ permet d'obtenir $f(n)_{\emptyset\text{-}U}$

En effet, soit k le nombre de référentiels de R(U):

$$f(n)_{\emptyset\text{-}U} = \sum_{(n_1,\ldots,n_k)/\sum_{i=1}^{k} n_i = n} \left(\prod_{i=1}^{k} f(n_i)_{\emptyset\text{-}U|\bar{R}} \right)$$

$$\forall i \; n_i \leq |\bar{R}_i(U)|$$

2) $X \subsetneq U$

Ce cas sera traité lors de la démonstration de la validité de P" pour l'équi-join.

- La démonstration de la validité de P" pour la différence se fait de façon tout à fait similaire.

Théorème II-5 :

P" est valide pour l'opération d'union dans CE. □

Démonstration :

Soit l'opération $R"(U) = R(U) \cup R'(U)$

Il faut déterminer $f"(n)_{\emptyset\text{-}X}$, $\forall X \subseteq U$

Soit $\bar{R}"(U)$ un référentiel de $R"(U)$ construit comme lors de la démonstration du théorème I-3 du chapitre IV.

Nous supposons que $\bar{R}"(U) \subset (U\bar{R}(U)) \cap (U\bar{R}'(U))$.

Les autres cas ($\bar{R}"(U) \cap (U\bar{R}(U)) = \emptyset$ et son symétrique) sont triviaux.

1) Nous regardons d'abord le cas où X = U

$$f(n)_{\emptyset\text{-}U|\bar{R}"(U)} = \sum_{i=1}^{n} \sum_{j=n-i}^{n} f(i)_{\emptyset\text{-}U} \times f'(j)_{\emptyset\text{-}U} \times \text{prob}(|R(U) \cap R'(U) \cap \bar{R}"(U)| = i+j-n \;/\; |R(U)| = \text{et } |R'(U)| =$$

Déterminer
prob($|R(U) \cap R'(U) \cap \bar{R}"(U)| = i+j-n \;/\; |R(U)| = i$ et $|R'(U)| = j$)
se fait de la même façon que lors de la démonstration du théorème précédent.

2) $X \subsetneq U$: $f"(n)_{\emptyset\text{-}X}$ s'obtient de la même façon.

$$f(n)_{\emptyset\text{-}X|\bar{R}"(U)[X]} = \sum_{i=1}^{n} \sum_{j=n-1}^{n} f(i)_{\emptyset\text{-}X} \times f'(j)_{\emptyset\text{-}X} \times \text{prob}(|R(X) \cap R'(X) \cap \bar{R}"(X)| = i+j-n \;/\; |R(X)| = i \text{ et } |R'(X)| = j)$$

La encore, la connaissance de $f"(n)_{\emptyset\text{-}X|\bar{R}"(U)[X]}$ pour tout référentiel de l'union permet d'obtenir $f"(n)_{\emptyset\text{-}X}$

Corollaire II-1 :

P" est valide pour l'opération de sélection dans CE. □

Théorème II-6 :

P" est valide dans CE pour l'opération d'équi-join. □

Démonstration :

$R''(XYZ) = R(XY)\ [X = X]\ R'(XZ)$

$\forall\ V \subseteq XYZ$ il faut déterminer $f''_{\emptyset\text{-}V}(n)$

1) $X \subseteq V$; soit $V = XY'Z'$

$$f''_{\emptyset\text{-}V}(k) = \sum_{i=1}^{k} f''_{\emptyset\text{-}X}(i) \times \sum_{\substack{(n_1,\ldots,n_i)\\(n'_1,\ldots,n'_i)}} \Big/ \sum_{j=1}^{i} n_j n'_j = k \left(\prod_{j=1}^{i} f_{X\text{-}Y'}(n_i) \times f'_{X\text{-}Z'}(n'_i) \right)$$

Nous avons vu lors de l'étude de l'opération d'intersection que nous savions déterminer $f''_{\emptyset\text{-}X}(i)$ (nous rappelons que $R''(XYZ)[X] = (R(XY) \cap R(XZ))\ [X]$), $f_{X\text{-}Y'}(i)$ et $f'_{X\text{-}Z'}(i)$ sont déterminées par P" (Théorème III-1).

2) $X \subsetneq V$; $V = X'Y'Z'$

$$f''_{\emptyset\text{-}V}(k) = \sum_{i=1}^{\infty} f''_{\emptyset\text{-}X}(i) \times \left[\sum_{\substack{(n_1,\ldots,n_i)\\(n'_1,\ldots,n'_i)}} \Big/ \sum_{j=1}^{i} n_j n'_j \geq k \left(\prod_{j=1}^{i} f_{X\text{-}Y'}(n_j) \times f'_{X\text{-}Z'}(n'_j) \right) \right.$$

$$\left. \times\ \text{prob}\left(\left| \bigcup_{j=1}^{i} R''(x_j, X'Y'Z') \right| = k \ /\ |R''(x_j, X'Y'Z')| = n_j n'_j \right) \right]$$

Comme nous l'avions indiqué lors de la démonstration de la validité de P" pour l'intersection, le cas $f''_{\emptyset\text{-}V}(n)$ pour $V \subsetneq X$ est un cas particulier de 2).

D'après le théorème de Jordan :

$$\text{prob}(|\bigcup_{j=1}^{i} R''(x_j,X'Y'Z')| = k/|R''(x_j,X'Y'Z')| = n_j n_j') =$$

$$(*)\quad \text{prob}([\sum_{r=1}^{i} (-1)^{r-1} \sum_{1 \le j_1 < j_2 .. < j_r \le i} |R''(x_{j_1},X'Y'Z') \cap .. \cap R''(x_{j_r},X'Y'Z')|] =$$

$$k/|R''(x_j,X'Y'Z')| = n_j n_j')$$

Nous avons vu que notre système de paramètres permet d'évaluer $P(|R \cap R'| = k/|R|=i$ et $|R'|=j)$. Nous pouvons donc évaluer à partir de ce système la probabilité (*).

Théorème II-7 :

P" est valide pour le join généralisé dans CE. □

Démonstration :

Elle est identique à celle du théorème I-9. Nous savons (d'après le corollaire III-1) que P" est valide pour les sélections générales . Il est clair que P" est valide pour le produit cartésien.

En effet : soit $R''(UV) = R(U) \times R'(V)$, $\forall\ X \subseteq UV$

$$f''_{\emptyset\text{-}X}(n) = \sum_{i,j/ij=n} f(i)_{\emptyset\text{-}X\cap U} \times f''_{\emptyset\text{-}X\cap V}(j)$$

$$(\text{Nous posons } f(i)_{\emptyset\text{-}X} = \begin{cases} 0 \ \forall\ i \neq 1 \\ 1 \text{ si } i = 1 \end{cases}).$$

P" est valide pour le join généralisé puisque celui-ci est une requête comportant un produit cartésien et une sélection (Théorème VI-3 du chapitre I).

Nous avons donc là un système de paramètres permettant d'évaluer la distribution de taille des relations résultats d'une requête formulée à l'aide de l'algèbre relationnelle. Cependant, pour des requêtes faisant appel à des opérations binaires, (union, intersection, join), l'évaluation du résultat peut se révéler difficile, principalement si la requête possède un degré de complexité nécessitant un grand nombre de référentiels pour chaque relation opérande afin de représenter les différentes propriétés de dépendance ou d'indépendance des projections de ces relations entre elles.

CHAPITRE VI

- Validité du système P' -

I - Définitions (rappel)

II - Validité du système P' dans la classe CI

Dans ce chapitre, nous allons montrer que le système P' (défini dans le chapitre II) est valide dans la classe CI (définie aussi dans le chapitre II). Le résultat ne conduit pas à une évaluation pratique des distributionsde taille des relations résultats d'une requête. En effet, le système P' est défini sur une classe assez générale, qui, en particulier, ne comporte aucune hypothèse simplificatrice sur la probabilité d'un n-uplet donné d'apparaître dans une valeur de relation d'un schéma de la classe. Cependant, la validité de ce système qui est,ainsi que la classe CI , une généralisation des systèmes et de la classe précédents, renforce la valeur théorique de la notion de schéma probabilisé en donnant une vue plus large des implications du modèle que nous avons construit.

I - DÉFINITIONS (RAPPEL)

La définition de la classe CI a été donnée en I.9 dans le Chapitre II. Nous rappelons la définition du système P'.

Définition I-1

Nous notons P' le système de paramètres suivant :

Soit (S,p) un schéma probabilisé, P' associe à (S,p)

- une valeur de REF
- $\forall R_t(U) \in R_t$, $\forall X \subseteq U$, $\forall X \in U\bar{R}(X)$: $f^{(i)}_{X=X-Y}$. □

Nous rappelons que ce système donne, pour chaque valeur des référentiels associés au schéma S, la probabilité de son appartenance à une projection d'une valeur de relation du schéma. Pour plus de précision, il faudrait indicer par le nom de la relation à la projection de laquelle peut appartenir l'élément du référentiel la distribution $f^{(i)}_{X=x-X}$.

Nous ne le ferons pas car cela alourdirait considérablement les notations alors qu'il n'y a pas de confusion possible.

II - VALIDITÉ DU SYSTÈME P' DANS LA CLASSE CI

Théorème II-1:

P' est valide dans CNE pour l'opération de projection. □

Démonstration :

1) Nous démontrons d'abord que les paramètres se "transfèrent" à travers l'opération de projection.
Soit : $(S,p) \in CNE$, $R(U) \in \mathcal{R}$. $R'(X) = R(U)[X]$ où $X \subset U$.

Il faut déterminer, $\forall Y \subseteq X$: $f'^{(i)}_{Y=y-Y}$ $\forall y \in \cup\bar{R}(Y)$.

Nous savons que $\forall \bar{R}(Y)$, $\bar{R}'(Y) = \bar{R}(Y)$. Il est clair que :
$\forall y \in \bar{R}(Y)$ $f'^{(i)}_{Y=y-Y} = f^{(i)}_{Y=y-Y}$.

2) D'après la définition de la validité, il faut montrer que P' permet de calculer les distributions de taille des projections des relations de la base.

$(S,p) \in C$, $R(U) \in S$ et $X \subseteq U$.

Nous allons d'abord nous placer dans le cas 1 de la propriété 2 (indépendance intra-relationnelle) : $\forall (x,x') \in \cup\bar{R}(X)$: $(x \in R(X))$ est indépendant de $(x' \in R(X))$.

Il faut déterminer $f^{(n)}_{\phi-X}$. Posons :

$$\gamma = |\cup\bar{R}(X)|$$

et $S(\gamma)$ le groupe des permutations d'un ensemble de γ éléments. Nous avons :

$$f^{(n)}_{\phi-X} = \sum_{\sigma \in S(\gamma)} \text{Prob}\,[(x_{\sigma(1)} \text{ et...et } x_{\sigma(n)} \in R(X)) \text{ et } (x_{\sigma(n+1)} \text{ et...et } x_{\sigma(\gamma)} \notin R(X))$$

D'après la propriété 2 , nous pouvons écrire :

$$\text{Prob}((x_{\sigma(1)} \text{ et...et } x_{\sigma(n)} \in R(X)) \text{ et } (x_{\sigma(n+1)} \text{ et...et } x_{\sigma(\gamma)} \notin R(X))) =$$

$$\text{Prob}((x_{\sigma(1)} \text{ et... et } x_{\sigma(n)} \in R(X)) \times \text{Prob}((x_{\sigma(n+1)} \text{ et... et } x_{\sigma(\gamma)}) \notin R(X))$$

Toujours d'après l'indépendance intra-relationnelle, nous pouvons écrire :

$$\text{Prob}(x_{\sigma(1)} \text{ et...et } x_{\sigma(n)}) \in R(X)) = \text{Prob}(x_{\sigma(1)} \in R(X)) \times \ldots \times \text{Prob}(x_{\sigma(n)} \in R(x))$$

et :

$$\text{Prob}(x_{\sigma(n+1)} \text{ et...et } x_{\sigma(\gamma)}) \notin R(X)) = \text{Prob}(x_{\sigma(n+1)} \notin R(X)) \times \ldots \times \text{Prob}(x_{\sigma(\gamma)} \notin R(X))$$

d'où puisque : $\text{Prob}(x \in R(X)) = f^{(1)}_{X=x-X}$ et $\text{Prob}(x \notin R(X)) = f^{(0)}_{X=x-X}$

$$f^{(n)}_{\phi-X} = \sum_{\sigma \in S(\gamma)} \left[\prod_{i=1}^{n} f^{(1)}_{X=x_{\sigma(i)}-X} \times \prod_{i=n+1}^{f} f^{(0)}_{X=x_{\sigma(i)}-X} \right]$$

Supposons maintenant que nous soyons dans l'autre cas de la propriété 2.

$\exists (V,W)$ une partition de X telle que $V \rightarrow W$ dans R(X). Ici nous n'avons plus indépendance.

En effet, par exemple vw et vw' ne peuvent exister dans la même relation (définition des dépendances fonctionnelles).

V est une clef de X donc : $|R(X)| = |R(V)|$, soit :

$$f^{(n)}_{\phi-X} = f^{(n)}_{\phi-V}$$

Il faut déterminer $f^{(n)}_{\phi-V}$.

$$v \in R(V) \Rightarrow \exists w \in \cup\bar{R}(W) \; / \; vw \in R(VW)$$

(Nous rappelons que VW = X).

Posons $|\cup\bar{R}(V)| = \gamma'$ et W^n une liste de k éléments de $\cup\bar{R}(W)$ distincts ou non. Alors nous avons :

$$f^{(n)}_{\phi-V} = \sum_{\sigma \in S(\gamma)} \left[\sum_{W^n} \prod_{i=1}^{k} f^{(1)}_{X=V^{w_i}_{\sigma(i)}-X} \right] \times \left[\sum_{W^{\gamma'-n}} \prod_{i=k+1}^{\gamma'} f^{(0)}_{X=V^{w_i}_{\sigma(i)}\, y-X} \right]$$

Ici nous avons bien le droit de considérer $f^{(1)}_{X=V_i^w-X}$ et $f^{(1)}_{X=V_j^{w'}-X}$ comme des probabilités indépendantes puisque V_i est différent de V_j, donc $\forall w$ et $w' \in W^n$ ou $W^{\gamma'-n}$, $V_i w$ et $V_j w'$ peuvent appartenir à la même relation (lemme I.1 du chapitre II).

Dans la formule précédente w_i varie sur les éléments de la liste W^n ou $W^{\gamma'-n}$ et ce pour toutes les listes possibles.

Nous voyons donc que dans les deux cas de la propriété 2, le système P' permet de déterminer la distribution de taille des projections des relations de la base.

Théorème II-2

P' est valide pour l'opération de sélection élémentaire sur l'égalité $(R(U)|X = x)$. □

- $(R(U)|X = x) = R'(U)$.

$\forall X' \subseteq U$, il faut déterminer : $\forall x' \in \cup\bar{R}(X') : f'^{(1)}_{X'=x'-X'} = \text{Prob}(x' \in R'(X))$.

Nous rappelons que : $\forall \bar{R}(U)$ un référentiel de $R(U)$:

$\bar{R}'(U) = \bar{R}(U)|X = x$.

1er Cas : $X \cap X' = \emptyset$

$$f'^{(1)}_{X'=x'-X'} = f^{(1)}_{X'X=xx'-X'}$$

En effet, x' n'apparaît dans R'(U) que si x'x apparaît dans R(U)[XX'].

2ème cas : $X' = X$

$$f'^{(1)}_{X=x'-X'} = \begin{cases} 0 \text{ si } x' \neq x \\ 1 \text{ si } x' = x \end{cases}$$

3ème cas : $X' \subset X$

$$f'^{(1)}_{X'=x'-X'} = \begin{cases} 0 \text{ si } x' \text{ n'apparaît pas dans } x \\ 1 \text{ sinon.} \end{cases}$$

4ème cas : $X \subset X'$

$$f'^{(1)}_{X'=x'-X} = \begin{cases} 0 \text{ si } x \text{ n'apparaît pas dans } x' \\ f^{(1)}_{X'=x'-X} \times \dfrac{1}{f^{(1)}_{X=x-X}} \end{cases}$$

5ème cas : $X \cap X' \neq \emptyset$ et $X \not\subseteq X'$ et $X' \not\subseteq X$

$$f'^{(1)}_{X'=x'-X'} = \begin{cases} o \text{ si } x'(X \cap X') \neq x(X \cap X') \\ f^{(1)}_{X'=x'-X} \times \dfrac{1}{f^{(1)}_{X\cap X'=x'(X\cap X') - X\cap X'}} \end{cases}$$

Dans les deux derniers, le deuxième terme dans l'accolade désigne la probabilité que x' soit dans R'(U)[X'] conditionnée par la probabilité que $X \cap X'$ est dans R'(U) $[X\cap X']$.

Nous allons montrer que pour cette opération, P permet de déterminer facilement les distributions de tailles des projections de R'(U) sans avoir à calculer les distributions $f'^{(1)}_{X'=x'-X'}$.

$\forall X'Y \subseteq U$, il faut déterminer : $f'^{(n)}_{X'=x'-Y} \quad \forall n$.

1er cas : $X \cap X' = \emptyset$

$$f'^{(n)}_{X'=x'-Y} = f^{(n)}_{XX'=xx'-Y}.$$

En effet, x' n'apparait dans R'(U) que si xx' apparaît dans R(U).

2ème cas : $X' = X$

$$f'^{(n)}_{X=x'-Y} = \begin{cases} 0, \forall n \text{ si } x' \neq x \\ f^{(n)}_{X=x-Y} \quad \text{sinon} \end{cases}$$

3ème cas : $X' \subset X$

$$f'^{(n)}_{X'=x'-Y} = \begin{cases} 0, \forall n \text{ si } x' \text{ n'apparaît pas dans } x \\ f^{(n)}_{X'=x'-Y} \quad \text{sinon} \end{cases}$$

4ème cas : $X' \supset X$

$$f'^{(n)}_{X'=x'-Y} = \begin{cases} 0, \forall n \text{ si } x \text{ n'apparaît pas dans } x' \\ f^{(n)}_{X'=x'-Y} \quad \text{sinon} \end{cases}$$

$\forall \bar{R}(U)$ un référentiel de $R(U)$, le référentiel correspondant de $R'(U)$ est $\bar{R}'(U) = \bar{R}(U)|X = x$.

Il faut maintenant montrer que P' permet de déterminer $f^{(n)}_{X=x-Y}$ $\forall (X,Y) \subseteq U$.

Soit $\gamma = |\cup\bar{R}(Y)|$ et $S(\gamma)$ l'ensemble des permutations sur un ensemble de γ éléments.

$$f^{(n)}_{X=x-Y} = \sum_{\sigma \in S(\gamma)} [\text{Prob}((xy_{\sigma(1)} \text{ et...et } xy_{\sigma(n)}) \in R(XY) \text{ et } (xy_{\sigma(n+1)} \cdots xy_{\sigma(\gamma)}) \notin R(XY)]$$

1er cas : $X \to Y$

Dans ce cas les probabilités sont :

$\forall n > 1$

$$\text{Prob}((xy_{\sigma(1)} \text{ et...et } xy_{\sigma(n)}) \in R(XY)) = 0$$

et $$\text{Prob}(xy \in R(XY)) = f^{(1)}_{XY=xy-XY}$$

$$\text{Prob}(xy \notin R(XY)) = f^{(0)}_{XY=xy-XY}$$

d'où :

$$f^{(n)}_{X=x-Y} = 0 \quad \forall n > 1.$$

$$f^{(1)}_{X=x-Y} = \sum_{y \in \cup\bar{R}(Y)} f^{(1)}_{XY=xy-XY}$$

et $$f^{(0)}_{X=x-Y} = \sum_{y \in \cup\bar{R}(Y)} f^{(0)}_{XY=xy-XY} = f^{(0)}_{X=x-X}$$

Théorème II-3

P' est valide pour l'opération d'intersection et de différence. □

Démonstration :

Il faut déterminer les distributions associées à la relation résultat d'une intersection.

Soit : $R(U) \cap R'(U) = R''(U)$

Soient : $\bar{R}(U)$ et $\bar{R}'(U)$ deux référentiels quelconques de $R(U)$ et $R'(U)$.

Nous nous plaçons dans le premier cas de la propriété 2 :

$\forall u, u' \in \bar{R}(U)$: $u \in R(U)$ est indépendant de $u' \in R(U)$

et

$\forall u, u' \in \bar{R}'(U)$: $u \in R'(U)$ est indépendant de $u' \in R'(U)$

soit :

$$\bar{R}''(U) = \bar{R}(U) \cap \bar{R}'(U)$$

et :

$$|\bar{R}''(U)| = \gamma.$$

Nous allons d'abord déterminer : $f^{(1)}_{U=u\,-U}$.

Nous nous plaçons d'abord dans le premier cas de la propriété 1 (indépendance inter-relationnelle).

Soit : $\forall u \in \bar{R}(U) \cap \bar{R}'(U)$, $u \in R(U)$ est indépendant de $u \in R'(u)$, dans ce cas, nous avons : $\forall u \in \bar{R}''(U)$:

$$f''^{(1)}_{U=u\,-U} = f^{(1)}_{U=u\,-U} \times f''^{(1)}_{U=u\,-U}$$

Dans le deuxième cas de la propriété 1 : $R(U) \cap \bar{R}'(U) \subseteq R'(U) \cap \bar{R}(U)$, nous avons : $\forall u \in \bar{R}''(U)$

$$f''^{(1)}_{U=u\,-U} = f^{(1)}_{U=u\,-U}$$

Le troisième cas est symétrique du second.

Nous allons maintenant déterminer : $f''^{(k)}_{\emptyset-U/\bar{R}''(U)} = \text{Prob}\{|R''(U)\cap\bar{R}''(U)| = k\}$.

Soit $S(\gamma)$ le groupe des permutations de γ éléments (nous rappelons que $|\bar{R}''(U)| = \gamma$), nous avons :

$$f^{2(k)}_{\emptyset-U/\bar{R}''(U)} = \sum_{\sigma\in S(\gamma)} \left(\prod_{i=1}^{k} f''^{(1)}_{U=u_{\sigma(i)}-U}\right) \times \prod_{j=k+1}^{\gamma} \left(1 - f''^{(1)}_{U=u_{\sigma(i)}-U}\right)$$

Il ne reste plus qu'à remplacer $f''^{(1)}_{U=u-U}$ par son expression en fonction des distributions de R'(U) et R(U).

Nous allons maintenant faire le même calcul dans le second cas de la propriété 2 :

$\forall$ u, u' $\in \bar{R}(U)$, u $\in$ R(U) et u' $\in$ R(U) si et seulement si u et u' ne sont pas incompatibles. Soit, d'après la définition de l'incompatibilité donnée dans le chapitre II (paragraphe I, lemme I-1 et définition I-1) :
$\exists(X,Y)$ une partition de U, telle que $(X\rightarrow Y)$ est vraie dans R(U).

Ici nous n'avons plus indépendance entre $f_{U=u-U}$ et $f_{U=u'-U}$.

En effet : $f^{(1)}_{U=u'-U} = 0$ si u $\in$ R(U).

Pour terminer, il faut déterminer $f''^{(n)}_{X=x-X}$ $\forall X \subset U$. Posons : $Y = U \setminus X$ et $|\bar{R}''(Y)| = \gamma$.

$$f''^{(1)}_{X=x-X} = \text{Prob}[(xy_1\in R(U)\cap R'(U)) \text{ ou } (xy_2\in R(U)\cap R'(U))\ldots\text{ou } (xy_\gamma\in R(U)\cap R'(u))]$$

Posons :

$$f''^{(1)}_{U=xy-U} = \text{Prob}\ (xy_1 \in R(U) \cap R'(u))$$

(nous avons vu que nous savons déterminer $f''^{(1)}_{U=u-U}$, d'où d'après les résultats connus sur la probabilité d'une union d'évènements :

$$f''^{(1)}_{X=x-X} = \sum_{i=1}^{\gamma} f''^{(1)}_{U=xy_i-U} - \sum_{i=1}^{\gamma}\sum_{j=i}^{\gamma} (f''^{(1)}_{U=xy_i-U} \times f''^{(1)}_{U=xy_j-U})$$

$$+ \sum_{i=1}^{\gamma}\sum_{j=i}^{\gamma}\sum_{k=j}^{\gamma} (f''^{(1)}_{U=xy_i-U} \times f''^{(1)}_{U=xy_j-U} \times f''^{(1)}_{U=xy_k-U}) - \ldots$$

Le cas de la différence se traite de façon similaire.

Théorème II-4

P' est valide pour l'opération d'union. □

Remarque : Nous rappelons que les référentiels de l'union sont construits comme indiqué dans la démonstration de la fermeture de CE pour l'union dans le chapitre IV.

Démonstration :

Soit : $R''(U) = R(U) \cup R'(U)$.

Soit $\bar{R}''(U)$ un référentiel de $R(U)$ obtenu à partir de deux référentiels $\bar{R}'(U)$ et $\bar{R}(U)$ comme indiqué dans la remarque. Il y a donc trois cas possibles :

$\bar{R}''(U) = \bar{R}(U) \cap \bar{R}'(U)$

$\bar{R}''(U) = \bar{R}(U)$

$\bar{R}''(U) = \bar{R}'(U)$

1er cas : $\bar{R}''(U) = \bar{R}(U) \cap \bar{R}'(U)$

1) Nous allons d'abord supposer que nous sommes dans le premier cas de la propriété 2 :

$\forall u,u' \in \bar{R}(U)$: $u \in R(U)$ est indépendant de $u' \in R(U)$ et
$\forall u,u' \in \bar{R}'(U)$: $u \in R'(U)$ est indépendant de $u' \in R'(U)$

Nous allons d'abord déterminer $f''^{(1)}_{U=u-U}$. Nous nous plaçons dans le premier cas de la propriété 1 : soit $\forall u \in \bar{R}''(U)$, $u \in R(U)$ est indépendant de $u \in R'(U)$. Dans ce cas, nous avons :

$\forall u \in \bar{R}''(U)$

$$f''^{(1)}_{U=u-U} = f^{(1)}_{U=u-U} + f'^{(1)}_{U=u-U} - \left(f^{(1)}_{U=u-U} \times f'^{(1)}_{U=u-U}\right)$$

Dans le deuxième cas de la propriété 1 : $\bar{R}'(U) \cap \bar{R}''(U) \subseteq R(U) \cap \bar{R}''(U)$, nous avons :

$$f''^{(1)}_{U=u-U/\bar{R}''(U)} = f'^{(1)}_{U=u-U/\bar{R}''(U)}$$

Le troisième cas est symétrique du second.

Pour des raisons de simplification, nous posons les distributions : $f^{(1)}_{U=u-U} = 0$ (resp. $f'^{(1)}_{U=u-U} = 0$) pour u n'appartenant pas au référentiel $\bar{R}(U)$ (resp. $\bar{R}'(U)$).

2) Nous nous plaçons dans le deuxième cas de la propriété deux : $\exists (X,Y)$ une partition de U telle que $X \rightarrow Y$ dans R(U) et dans R'(U).

Dans ce cas $X \rightarrow Y$ dans R"(U) et R"(U) vérifie le deuxième cas de la propriété deux ($f''^{(i)}_{U=u-U}$ n'est pas changé).

3) Le dernier cas est : $\exists (X,Y)$ une partition de U telle que $X \rightarrow Y$ dans R(U) et $X \not\rightarrow Y$ dans R'(U).

Dans ce cas $X \not\rightarrow Y$ dans R"(U) et R"(U) vérifie le premier cas de la condition deux.

Pour terminer le 1er cas, il faut calculer $f''^{(n)}_{X=x-X}$ $\forall X \subset U$.

Par définition de l'union : $x \in (R(U) \cup R'(U)[X]$ si $x \in R(U)[X]$ ou si $x \in R'(U)[X]$, donc :

$$f''^{(1)}_{X=x-X} = f^{(1)}_{X=x-X} + f'^{(1)}_{X=x-X} - f''^{(1)}_{X=x-X}$$

où $f''^{(1)}_{X=x-X}$ est la probabilité que $x \in (R(U) \cap R'(U))[X]$.

Nous avons vu lors de la démonstration du théorème II.3 que nous savons déterminer $f''^{(1)}_{X=x-X}$.

2ème cas : $\bar{R}''(U) = \bar{R}(U)$

D'après la construction des référentiels donnée dans la démonstration du théorème I-3 du chapitre IV, dans ce cas $\bar{R}(U) \cap (\cup\bar{R}'(U)) = \emptyset$. Donc :

$$(R(U) \cup R'(U)) \cap \bar{R}''(U) = R(U) \cap \bar{R}''(U)$$

Donc : $\forall u \in \bar{R}''(U)$:

$$f''^{(1)}_{U=u-U} = f^{(1)}_{U=u-U}$$

Remarque : $f''^{(k)}_{\emptyset-U/\bar{R}''(U)} = f^{(k)}_{\emptyset-U/\bar{R}''(U)}$.

3ème cas : $\bar{R}''(U) = \bar{R}'(U)$: symétrique du cas précédent.

Corollaire II-1

Le système P' est valide pour l'opération de sélection dans CNE. □

Nous rappelons que ce corollaire se déduit des Lemmes I-1 et I-2 du chapitre IV et des théorèmes II-2, II-3, II-4.

Théorème II-5

P' est valide pour le join généralisé. □

Démonstration :

La démonstration découle directement du lemme I-3 du chapitre IV.

En effet, il suffit de prouver que P' est valide sous $\mathcal{E}$ pour l'opération de sélection et pour le produit cartésien.

En effet, soient R(X) et R'(Y) deux relations :

$$\forall X'Y' \subseteq XY \text{ et } (x'y') \in (\cup\bar{R}(X)[X']) \times (\cup\bar{R}'(Y)[Y'])$$

Si $R''(XY) = R(X) \times R'(Y)$, alors :

$$f''^{(1)}_{X'Y'=x'y'-XY} = f^{(1)}_{X'=x'-X'} \times f''^{(1)}_{Y'=y'-Y'}$$

P' est donc valide pour cette opération, ce qui termine la démonstration.

□

CONCLUSION

Dans ce qui précède nous avons présenté un modèle formel permettant l'étude de différents moyens d'évaluation des tailles des relations construites lors du traitement d'une requête. De ce modèle, nous avons obtenu un certain nombre de résultats théoriques ou pratiques d'intérêts variés. Nous allons souligner ici les principaux points importants de ce travail. Nous pourrons dégager deux principales contributions : le modèle formel et les applications.

1) Modèle formel :

Nous avons donné au problème de l'évaluation de la taille des requêtes une structure probabiliste bien définie présentant de nombreux avantages tant pour l'étude théorique du comportement d'une base de données que pour l'obtention de résultats pratiques applicables dans les SGBD. Ce modèle nous a permis de définir de façon rigoureuse la nature des paramètres sur la base de données que nous utilisons pour l'évaluation des relations dérivées.

De même nous avons défini les propriétés générales que ces paramètres doivent vérifier afin de permettre l'évaluation de la taille de toute requête construite à partir de l'algèbre relationnelle définie par Codd [Codd 71].

2) Applications

Nous avons défini deux classes de bases de données dans le cadre desquelles nous avons montré qu'il existait des ensembles de paramètres permettant d'évaluer soit l'espérance de la taille des relations dérivées, soit la distribution de la taille et ce pour

toute requête construite à partir de l'algèbre relationnelle.

De plus, nous avons donné les résultats d'évaluation obtenus pour chaque opération de l'algèbre relationnelle dans le cadre d'une de ces classes. Nous avons ainsi obtenu l'espérance de la taille des relations construites à partir des relations d'une base de données de cette classe pour les opérateurs de projection, sélection, intersection, union et join. De plus nous avons étudié et mis en évidence les liens existants entre les principales contraintes d'intégrité sur les bases de données relationnelles et notre modèle. Ainsi nous avons montré que la connaissance de ces contraintes permet de simplifier les ensembles de paramètres nécessaires pour l'évaluation et les calculs impliqués par celle-ci. La raison d'être de ce travail est de permettre la création d'algorithmes d'optimisation plus fins et plus généraux tenant compte de la taille des relations dérivées construites lors du traitement d'une requête. En effet, il n'existait pas jusqu'ici de moyens précis de décider dans quel ordre traiter un join multiple par exemple, ou dans le cas d'une base de données répartie de choisir les sites sur lesquels effectuer les opérations constituant une requête. Dans ce cadre, les résultats d'évaluation de l'espérance de la taille des relations dérivées sont directement utilisables.

RÉFÉRENCES

[Aho 79] Aho A., Sagiv Y., and Ullman J.D., "Equivalences array for relational expressions", SIAM J. Comput., Vol. 8, N° 2; May 1979.

[Arms 74] Armstrong W.W., "Dependency structures of data base relationship", Information processing 74, North Holland Publishing Company (1974).

[Astr75a] Astrahan M.M. and Chamberlin D.D., "Implementation of a structured English query language", Commun. ACM 18, 10 (Oct. 1975), 580-588.

[Astr75b] Astrahan M.M. and Lorie R.A., "SEQUEL-XRM, a relational system", in Proc. ACM Pacific Regional Conf. April 1975, pp. 34-38.

[Beer 77] Beeri C., Fagin R., Howard J.H., "A Complete Axiomatization for Functional and Multivalued Dependencies in Database Relations", IBM Research Division, RJ 1977 (27927) 4/1/77, Comp. Sc.

[Beer 79] Beeri C., Mandelzon A.O., Sagiv Y., Ullman J.D., "Equivalence of relational database schemes", 1979 ACM.

[Blas 76] Blasgen M.W. and Eswaran K.P., "On the evaluation of queries in a relational data base system," Research report RJ1745, IBM Research, San Jose, Calif. April 76.

[Blas 77] Blasgen M.W. and Eswaran K.P., "Storage and access in relational data bases", IBM Syst. J. 16, 4 (1977).

[Chan 76] Chandra A.K., Merlin P.M., "Optimal implementation of conjunctive queries in relational Data Bases", Proc. 9th, Annual ACM Symp. on theory of computing 1976, pp. 77-90.

[Chang76] Chang P.Y., "Parallel processing and data driven implementation of a relational data base systems", in Proc. ACM Annual Conf. 1976, pp. 314-318.

[Chiu 80] Chiu D.M., Ho Y.C., "A Methodology for Interpreting Tree Queries into Optimal Semi-join Expression". Proc. of ACM-SIGMOD 1980, Int. Conf. on Management of Data May 14-16.

[Clau 80] Clausen S.E., "Optimizing the evaluation of calculus expressions in a relational database system", Inform. Systems Vol. 5, pp. 41-54, 1980.

[Codd 70] Codd E.F., "A relational model for large shared data banks", Commun. ACM 13, 6 (June 1970), 377-387.

[Codd71a] Codd E.F., "A data base sublanguage founded on the relational calculus", in ACM SIG FIDET Workshop on Data Description, Access and Control, 1971, pp. 35-68.

[Codd71b] Codd E.F., "Further normalization of the data base relational model", in Data Base Systems, Courant Computer Science Symposia, Vol. 6, Prentice Hall, Englewood Cliffs, N.J. May 1971.

[Codd71c] Codd E.F., "Relational completeness of data base sublanguages", in Data Base Systems, Courant Computer Science Symposia, Vol. 6, Prentice-Hall, Englewood Cliffs, N.J., May 1971.

[Delo 78] Delobel C., "Normalization and hierarchical Dependencies in the Relational data Model", ACM Trans. on database Systems, Vol. 3, N° 3, Sept. 1978, pp. 201-222.

[Demo 77] Demolombe R., "A general semantic method for efficiently evaluating "and" operators in relational DBMS" O.N.E.R.A C.E.R.T. Internal report n° LBD 7715, Nov. 77.

[Demo 80] Demolombe R., "Estimation of the number of tuples satisfying a query expressed in predicate calculus language", O.N.E.R.A. C.E.R.T. VLDB Montréal Canada, Oct. 1-3, 1980.

[Demo 81] Demolombe R., "Evaluation des tailles des projections des relations", communication personnelle.

[Epst 79] Epstein R., Stonebraker M;, Wong E., "Distributed query processing in a relational data base system", ACM-SIGMOD 79.

[Furt 77] Furtado A., Kerschberg L., "An algebra of quotient relations"

[Hall74a] Hall P., and Todd S., "Factorization of algebraic expressions", Rep. UKSC-55 IBM UK Scientific Centre, Peterlee, England, April 1974.

[Hall74b] Hall P., "Common subexpression identification in general algebraic systems", Rep. UKSC-60, IBM UK Scientific Centre, Peterlee, England, Nov. 1974.

[Hall 75] Hall P., "Optimization of single relational expressions in a relational data base system", Rep. UKSC-76, IBM UK Scientific Centre, Peterlee, England, June 1975.

[Kim 79] Kim W., "Relational Database Systems", Computing Surveys, Vol. 11, N° 3, Sept. 79.

[Kim 80] Kim W., "A new way to compute the product and join of relations", In proc of ACM-SIGMOD 1980. International Conference on Management of Data, May 14-16.

[Pale 72] Palermo F.P., "A data base search problem", Research Rep. TJ1072, IBM Research, San Jose, Calif, July 1972.

[Pech 75] Pecherer R.M., "Efficient evaluation of expressions in a relational algebra", in Proc. ACM Pacific Regional Conf. April 1975, pp. 44-49.

[Pech 76] Pecherer R.M., "Efficient exploration of product spaces", in Proc. Int. Conf. Management of Data (ACM), 1976, pp. 169-177.

[Rich 81] Richard Ph., "Evaluation of the size of a query expressed in relational algebra", to appear in Proc. of ACM-SIGMOD 1981, International Conference on Management of Data, April 29 - May 1.

[Roth 74] Rothnie J.B.Jr., "An approach to implementing a relational data management system", in 1974 ACM SIGMOD Workshop on Data Description, Access and Control, pp. 277-294.

[Roth 75] Rothnie J.B. Jr., "Evaluating interentry retrieval expressions in a relational data base management system", in Proc. 1975, AFIPS Nat. Computer Conf., Vol. 44, AFIPS Press., Montvale, N.J., pp. 417-423.

[Seli 79] Selinger P.G., Astrahan M.M., Chamberlin D.D., Lorie R.A., and Price T.G., "Access path selection in a relational database system", Research Rep. RJ2429, IBM Research, San Jose, Calif. Jan. 1979, ACM 1979 SIGMOD, pp. 23-34.

[Senk 73] Senko M.E., Altman E.B., Astrahan M.M. and Fehder P.L., "Data structures and accessing in data base systems", IBM Syst. J. 12, 1 (1973).

[Smit 75] Smith J.M. and Chang P.Y., "Optimizing the performance of a relational algebra database interface", Commun. ACM 18, 10 (Oct. 1975), 568-579.

[Spyr 77] Spyratos N., Bancilhon F., "Name Independence in relational data bases", Rapport de recherche n° 241, Août 77, INRIA

[Ston76a] Stonebraker M., Wong E., Kreps P. and Held G., "The design and implementation of INGRES", ACM Trans. Database Syst. 1, 3, (Sept. 1976), 189-222.

[Todd 74] Todd S., "Implementation of the join operator in relational data bases", Tech. Note 15, IBM UK Scientific Centre, Peterlee, England, Nov. 1974.

[Verh76a] Verhofstad J., "The PRTV optimiser : the current state", Rep. UKSC-83, IBM UK Scientific Centre, Peterlee, England, May 1976.

[Verh76b] Verhofstad J., and Todd S., "Cost prediction for database query optimization", Tech, Note 52, IBM UK Scientific Centre, Peterlee, England, Jan. 1976.

[Verh 77] Verhofstad J., "An evaluation of the PRTV optimiser", Tech., Note IBM UK Scientific Centre, Peterlee, England, 1977.

[Wong 76] Wong E., and Youssefi K., "Decomposition - a strategy for query processing", ACM Trans. Database Syst. 1, 3 (Sept. 76) 223-241.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique

I.S.B.N. 2 - 7261 - 0266 - 2